Workshop on Data Analysis WDA'2009

František Babič, Ján Paralič, Andreas Rauber (Eds.)

Proceedings of the
9th International Student Workshop

Čertovica (Low Tatras), Slovakia

July 2 – 4, 2009

Preface

The 9th International Student Workshop on Data Analysis (WDA) took place in Čertovica, a saddleback situated in the middle of the main ridge of the Low Tatras mountains. This year's program contained traditionally presentations of graduate and PhD. students from two different universities – Technical University in Košice and Vienna University of Technology.

The Workshop, held on July 2 - 4, 2009, consisted of four main thematic sessions. The first session was focused on music retrieval; the second one presented interesting topics from digital preservation domain. Third session dealt with various approaches to data and text analysis. The last session was oriented to projects based on service-oriented architectures.

The first session about music retrieval started with a paper presented by Andrei Grecu, who presented various algorithms and evaluation measures for musical instrument sound separation. The next contribution by Jakob Frank focussed on differences and similarities between user's playlist and its visual shape on a Music Map.

The next session stared with presentation of Stephan Strodl about the Hoppla archiving system that provides the possibility to archive digital objects in different formats in home user or smal office settings. In the second presentation Mark Guttenbrunner identified requirements for the selection of the best emulation tool for a set of digital objects. A Digital Preservation Time Capsule providing a showcase of the types and amount of representation information necessary to preserve digital objects was sketched by Natasha Surnic. The last contribution in this section was by Hannes Kulovits and described specific integration issues for recommender systems in the digital preservation planning tool Plato.

The third session consisted of three contributions dealing with specific data analysis and information extraction tasks. Gabriel Tutoky presented usability aspects of applying Named Entity Recognition in biomedical texts. The contribution by Martin Repka, described an experimental tool for the analysis of Citation Networks. the last presentation was given by Jozef Wagner, who proposed an analytical approach for the evaluation of a knowledge creation processes supported by a collaborative system based on monitored data stored as logs in a transactional database.

In last session, Zoltán Ďurčík presented his approach for automated web service composition and Karol Furdík identified key components of a software platform for secure, flexible, and project-oriented collaboration of business organisations within a temporary alliance.

We would like to thank all participants of the workshop for their contributions and very fruitful discussions after each presentation, making this year workshop an interesting and stimulating event.

October 2009 Editors

Content

Session 1: Music Retrieval

Session 2: Digital preservation

Session 3: Data and text analysis

Session 4: Service-oriented architecture

Session 1

Music retrieval

Challenges in Evaluating Musical Instrument Sound Separation Algorithms

Andrei Grecu

Vienna University of Technology, Vienna, Austria,
grecu@ifs.tuwien.ac.at,
WWW home page: http://ifs.tuwien.ac.at/~grecu/index.html

Abstract. In order to measure the separation quality of different algorithms automatically, corpora and error measures are needed. We argue that the few available corpora for blind audio source separation do not meet the quality criteria which we propose in this paper and so an urgent need for a better corpus exists. The lossy compression problematic is also discussed together with the necessary parameter specification of the filters used to generate the downmix. Finally we discuss the benefits of SNR, SIR, SAR and SDR as well as their shortcomings and also take a look at the problematic of optimally assigning estimated tracks to their corresponding originals in badly separated songs.

1 Introduction

Separating the sound of musical instruments from a mixture can be considered to be a special case of separating general audio signals. Separation in this context means segregating the sound of each instrument from a mixture, ideally resulting in separate tracks where each instrument performs solo without audible degradation or interference.

This work will concentrate on the evaluation of those algorithms, more specifically at the automatic evaluation without human intervention. For that purpose we need a good corpus or ideally more than one, a good error measure and a technique to measure the error of the separation algorithms on that corpus.

In Section 2 we will discuss some quality criteria and look at existing corpora, Section 3 deals with error measures and evaluation procedures and finally in Section 4 we will draw our conclusions.

2 Corpora

2.1 Quality Criteria

Corpora are a necessity for reasonable evaluation of different separation algorithms. The better the quality of the corpus the more meaningful the results will be. Unfortunately there is no actively maintained corpus with realistic reference tracks before mixdown and enough songs at the same time. Realistic here means that the tracks:

1. have to be full-length in order to give separation algorithms the possibility to adapt to the mixing statistics or other instrument information. 10-30 seconds is usually not enough for complex algorithms. We rather suggest 2:00 minutes.
2. should contain sounds of real instruments, at least this should be the majority of the tracks.
3. should have reasonable audio quality.
4. should have reasonable complexity, e.g. not a single continuous tone.

Furthermore the number of component tracks should be enough to make up a song with reasonable complexity and the entire corpus should contain enough songs in order to be able to create meaningful performance statistics. And an especially important point for the research community: the songs should have a license which permits redistribution for research purposes (i.e. creative commons) so the corpus can be made accessible to everyone interested.

Another aspect of quality is the kind of compression applied to the original sources, the mixdown and the results. We know lossless formats will are the best choice when it comes to quality, but how about lossy formats like MP3 and OGG? The problem of lossy formats lies in that they alter the original data. So if we lossily compress a file after the mixdown of the instruments, then perfect separation will not be possible even in theory due to the information lost in the compression. Therefore, in order preserve the possibility of perfectly separating instruments, at least in theory, only lossless audio compression should be use for downmixes. The same is true for the separated tracks. In order to not unnecessarily distort the error measures used in automatic evaluations those tracks should be stored in a lossless format, too, otherwise information will be lost in the separated track. On the other hand subjective evaluation by humans should not be affected by lossy compression of the resultant tracks. So until now we know that the mixture and the separated tracks have to be stored losslessly for automated evaluation. The question still remains whether the original tracks may be compressed with losses. A decisive answer for that question would serve well when constructing new corpora, as they consist only of original tracks and downmixes but we will not go deeper into that issue here, as it needs more investigations whether and how it affects separation performance. Still, we conjecture the following points:

- The inter-channel phase information, especially in masked out areas, may become too noisy to be usable. Masked out areas are those parts in the spectrum which can not be perceived by the brain due to a strong signal in the vicinity of that area.
- The high frequency content of a lossy compressed file may not have much in common with the original. This is especially true for advanced audio coding (AAC) based formats which will synthesize noise in the high frequency components shaped in such a way that it will sound like the original. However, that noise still has inter-channel phase information correlated with the originating instrument, but on the other hand it will not have harmonic information.

– Constant bitrate encoders have to vary the compression quality with time in order to not exceed their target bitrate. So if the compression affects separation, then the separation quality will then also vary with time. This is an effect which would be perceived as annoying by the human listener.

Actually, many artists use lossy compression for their component tracks in order to reduce the needed internet bandwidth to download them, so restricting a corpus to only contain original tracks with lossless compression may narrow down the number of candidate tracks too much.

2.2 Existing Corpora

After having introduced the quality criteria for a good corpus and discussing the choices of compression formats for the data, let us now look at some existing corpora and see in how far they meet our quality criteria:

– The BASS-dB [1] is a well selected corpus with 20 songs which is specialized on musical instrument separation or as the abbreviation suggest for blind audio source separation, but unfortunately it is not maintained anymore, so there are plenty of dead links. Only a few songs are still available with their original tracks. From those songs which are available some use either mp3 (lossy) or flac (lossless).
– Another specialized corpus for instrument separation is the IS Corpus [2] which is made of three parts: IS-B, IS-M and IS-R, each containing 4 songs. IS-B consists of binaural recordings made with microphones emulating a human head in order to preserve as many spatial cues as possible. Unfortunately IS-B has no component tracks at all so the performance cannot be evaluated automatically. IS-R has 4 realistic tracks meeting the quality criteria and are compressed losslessly but unfortunately 4 tracks are too few for a good statistic about separation performance. IS-M is made of module files (.MOD) which were pre-processed to have one instrument per track and then converted to WAVE. While these 4 tracks are fine quality-wise and are also stored using lossless compression, they may be considered unrealistic due to their nature of being computer generated music with files being kept small and simple so as to fit the demos and games that they were designed for. Besides IS-M also contains only 4 songs which are few.

2.3 Parameter and Usage Specification

So far we have discussed the most important issues for the data of the corpus but now we need some rules on how to use the corpus. For example mixing the original tracks down to a single file is not that straightforward as it may seem because there are some parameters which have to be agreed upon in order for everyone to get the same results:

- It is for example necessary to specify whether any delay is used for each track in part. The delay itself may be common for both channels which means that the concerned track will be played later in time, or only one channel may be delayed which translates to stereo panning. So the delay is a parameter which has to be agreed upon and specified together with the song. The specification of the delay is also important it has to be undone in order to compare the separated tracks with the originals. So it may prove useful to apply the delay to the originals and omit the delay parameter altogether.
- Another parameter which has to be either specified or incorporated in the original is the mixing gain for each track. The mixing gain is also important because of its influences in the weightings of the tracks in the error function. Neither specifying the mixing gain nor incorporating it in the sound files therefore may lead to different error values for the same original tracks.
- A common procedure for mixing real life songs is to add echoes to the mixture or to each track in part before mixing. So it has to be specified whether echoes are added to the mixture because algorithms are expected to remove the echoes once they are present.
- Further effects may be added after mixdown to simulate real music recordings and whose presence has to be listed in order for the algorithms to be know whether to try to remove them. Those are
  - dynamic compressors. This is a very common filter in popular music which applies more gain to silent parts of the music so that there is less difference between loud and silent parts. The desired effect is to make music be perceived louder or "hotter" and ultimately be bought by more people. Dynamic compressors are quite hard to remove but due to their widespread use one may decide to have some songs in the corpus compressed.
  - reverberation. Like echoes, this filter adds copies of the signal to itself, adding a more spatial feeling. Reverberation is usually very difficult to remove as it has many characteristics common to noise but as this filter is also widely used it should also be included in at least some tracks in order to make up a realistic corpus.
  - equalizer. An equalizer is a linear device and therefore separating algorithms have no possibility to guess its parameters, at least not without extensive knowledge of how the sounds contained in the tracks should be like. Therefore if an equalizer is used then it should be applied to the original tracks.

Some algorithms return normalized tracks, that is they either have all the same amount of energy or the same peak amplitude. Actually this should pose no problem as a simple least squares fitting inverse-gain coefficient can be calculated so that these tracks match the originals they belong to. We shall note that is not the same problem as not specifying the gain for each track before mixdown as this would lead to the algorithms erroneously giving importance to otherwise unimportant tracks. If the separation algorithm does not concentrate its efforts into separating the most energy-rich tracks once the gain parameters

were correctly given or applied, then this is not a problem of the corpus but of the algorithm itself.

3 Evaluation Measures and Procedures

3.1 Evaluation Measures

In order to get results from quality-wise ideal corpora, we need some plausible methods and measures to evaluate the separation algorithms. In this paper we will concentrate on automatic evaluation using error measures and leave out subjective evaluation. A popular measure is the signal to noise ratio (SNR) which is computed as

$$SNR = 10 \log_{10} \frac{\|\mathbf{x}\|^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2} \tag{1}$$

where $\mathbf{x}$ denotes the original track and $\hat{\mathbf{x}}$ the estimated signal. This measure has the advantage of being simple and easy to understand. Furthermore the SNR for multiple tracks and files can be calculated by simply concatenating them and calculating the SNR over the resulting vector. This comes especially handy if a single number is needed for the content of the whole corpus.

Another popular trio is the signal to interference ratio (SIR), signal to artefact ratio (SAR) and signal to distortion ratio (SDR), all of them which are described in [3]. All three measures are based on the following decomposition:

$$\begin{aligned} \hat{\mathbf{x}}_\mathbf{i} &= \mathbf{s}_\mathbf{i} + \lambda_\mathbf{i} + \epsilon_\mathbf{i} \\ \mathbf{s}_\mathbf{i} &= \alpha_i \mathbf{x}_\mathbf{i} \\ \lambda_\mathbf{i} &= \sum_{j \neq i}^{N} \beta_{i,j} \mathbf{x}_\mathbf{j} \end{aligned} \tag{2}$$

α_i and $\beta_{i,j}$ are coefficients chosen to best fit their original tracks $\mathbf{x_i}$ and $\mathbf{x_j}$ to the estimated signal $\hat{\mathbf{x}}_\mathbf{i}$, where i and j denote track numbers. $\mathbf{s_i}$ is called the target signal, $\lambda_\mathbf{i}$ the interference signal and $\epsilon_\mathbf{i}$ represents the remaining artefacts. From these three component signals we can now calculate the aforementioned ratios:

$$SIR = 10 \log_{10} \frac{\|\mathbf{s}_\mathbf{i}\|^2}{\|\lambda_\mathbf{i}\|^2} \tag{3}$$

$$SAR = 10 \log_{10} \frac{\|\mathbf{s}_\mathbf{i} + \lambda_\mathbf{i}\|^2}{\|\epsilon_\mathbf{i}\|^2} \tag{4}$$

$$SDR = 10 \log_{10} \frac{\|\mathbf{s}_\mathbf{i}\|^2}{\|\lambda_\mathbf{i} + \epsilon_\mathbf{i}\|^2} \tag{5}$$

Let us now look at some properties of these ratios. As we see, we now cannot simply concatenate vectors as we did for the SNR in order to obtain average Ratios due to the scaling coefficients α_i and β_j which have to be calculated for each track in part. But what we may do is concatenating the component signals $\mathbf{s_i}$

and $\lambda_\mathbf{i}$ and then calculate the ratios. So far so good, but let us now look at those scaling coefficients α_i and $\beta_{i,j}$. Their optimal calculation would imply solving an overdetermined linear system of equations by using least squares methods. While such an implementation may be fast we are confronted with the problem, that the resulting coefficients may also be negative which would lead to an invalid result as we do not expect to have to have the original signal flipped or to have interference multiplied by an negative factor. So in this case one would have to solve a linear system with constraints which needs much more computational effort than a simple least squares solution. A suboptimal algorithm can be chosen which computes each scaling factor in part but in that case the algorithm has to be specified in pseudocode because its results depend on the exact order of the different steps to the solution.

So, in summary while these three ratios are a good attempt for a better quantification of the different error components, their computational complexity is rather high or one has to resort to suboptimal solutions for the errors.

3.2 Evaluation Procedures

In order to calculate the errors on the separated tracks after separation has completed one first has to attribute them to their originals. This is usually accomplished by calculating the error measures for each combination of estimated and original track resulting in an error matrix and then chooses the combination with the smallest error for each track in part. As long as the tracks are well separated this heuristic will give optimal results but for not so well separated tracks this solution is suboptimal because the results depend on which track was first assigned to an original. So for badly separated tracks the results may also differ due to the assignment algorithm and not only by the separation capabilities of the algorithm which we want to measure. Even more, we suppose that the complexity of an optimal assignment is $O(NP)$. Fortunately the number of tracks is always small so using an optimal algorithm should not consume too many resources.

We noticed some practice for easing the separation task which aims to separate more than the target number of tracks and then add the surplus tracks to the others in a way to minimize the separation error. We do not see that as a fair practice if the number of target tracks is given to the separation algorithm beforehand because then one could split the mixture track into quite many components and then add them again together using the original tracks as reference which will then very probably lead to better results due to the available information about the originals than if just the target number of tracks was separated. So if more estimated tracks are generated than originals then they should be added together or discarded by the separation algorithm before evaluation.

4 Conclusions

We have discussed the challenging sides of evaluating musical instrument sound separation algorithms. The lack of corpora meeting the quality criteria we pro-

posed in this work signals an urgent problem which has to be tackled in the near future. We found that the compression format matters for the end results and in general lossless compression should be used for the downmix and end results in order to not unnecessarily distort the error figures. For the original tracks we could not come to a conclusive answer whether it will affect the separation quality negatively or not. We mentioned that the parameters and filters used for generating the downmix have to be specified beforehand. This is true for at least the delay and mixing gain parameters as the separation algorithms are not expected to undo them but may lead that spending more efforts on unimportant tracks. Furthermore four evaluation measures for quantifying the errors found in the separation results were discussed resulting that SNR is easier to compute while SIR, SAR, SDR need either many computational resources or have to be computed using suboptimal algorithms. On the other hand SIR, SAR, SDR do better express the different error components of the signal than SNR does. Assigning each estimated track to its original is argued to have a high complexity for badly separated sources because in that case the order of the assignments matters. We also argue that one should not use the original track information to postprocess tracks for example by adding them together after more tracks than originals were generated.

References

1. Vincent, E., Gribonval, R., Févotte, C., al.: *BASS-dB: the Blind Audio Source Separation Evaluation Database.* URL: `http://www.irisa.fr/metiss/BASS-dB/`
2. Grecu, A.: Musical Instrument Sound Separation: Extracting Instruments from Musical Performances - Theory and Algorithms. VDM Verlag Dr. Müller, Saarbrücken, Germany (2008)
3. Gribonval, R., Benaroya, L., Vincent, E., Févotte, C.: Proposals for performance measurement in source separation. In: Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003), Nara, Japan (2003) 763–768

Analysing and Evaluating Playlists on Music Maps

Jakob Frank

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/mir
frank@ifs.tuwien.ac.at

Abstract. Paths drawn to a *Music Map* are used to quickly and intuitively create playlists by drawing figures onto the map surface. In this paper the correlation between the quality of a playlist and its visual shape on a *Music Map* is investigated. A series of amateur and professional playlists together with generated playlists are analysed according to their visualisation on a *Music Map*. Furthermore, the playlists are evaluated in a small user study where rated quality by the users is compared to the graphical representation and the song distances in different feature spaces.

1 Introduction

One question that arises whenever someone wants to listens to music is: "Is it worth creating a playlist – or do I just press *shuffle*?" The former is getting more and more difficult, due to the sheer amount of music available on today's computers while the latter is rendered useless by the huge variety of music on any player. One approach to ease the playlist creation process is to provide the user an intuitive and interactive overview of his music collection, a *Music Map*, and a quick way to generate playlists.

The quality of the generated playlists is, however, often not satisfying. To improve the quality of the generated playlists, it is necessary to understand what determines the quality of a playlist. Since *Music Maps* create playlists based on figures drawn on the map, a series of playlists were analysed according to their visual shape. Furthermore, a small user study was lauched to gain more information about the correlation of the qualitiy of a playlist and its visual shape on a *Music Map*.

The remainder of this paper is structured as follows: Section 2 gives a brief overview about the technical background of *Music Maps*. In Section 3 the approach to visualise playlists is presented. A small scale user study about the correlation of the visual shape of a playlist and its quality is presented in Section 4. Finally, Section 5 summarises the conclusions and gives an outlook to future work.

2 Technical Fundamentals

The analysis and visualisation of playlists as it is presented in this paper relies on *Music Maps* which provide a graphical interface to large audio collections. [1]

The creation of such a *Music Map* is divided into two steps: feature extraction and map creation. First, the individual songs are analysed to extract descriptive features from the content of the audio stream. A wide variety of different feature extraction algorithms is available. For the experiments in this paper, Rhythm Patterns were used. The feature extraction process for a Rhythm Pattern is composed of two stages. First, the spectrogram of the audio is computed using the short time Fast Fourier Transfomr (STFT). After that, the Bark scale, and other psycho acoustic modells are applied to the spectrogram, aggregating it to 24 frequency bands. The spectrogram is then transformed to the Sone scale which reflects the human loudness sensation. In the second step, a discrete Fourier transform is applied to the Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each critical band. After further smoothing and weighting steps, Rhythm Patterns numerically represent magnitude of modulation for 60 modulation frequencies on 24 bands, thus resulting in a 1440-dimensional vector. High values for a specific modulation frequency in a number of adjacent bands indicate a specific rhythm occuring in a song. [3, 5]

After the feature extraction process, the resulting feature vectors are used as input for a self-organising map (SOM). A SOM is an neural network model that provides a projection from high dimensional data points to a lower, generally 2-dimensional output space. [2] A SOM iteratively arranges the data points in the output space is such a way, that data points which where located close to each other in the input space are also located nearby in the output space.

Applied to this domain, the SOM groups songs which share a common rhythm sturcture onto the same regions on the map. In detail, the algorithm works as follows. The map consists of a predefined number of units, which are arranged on a two-dimensional grid. Each of the units is assigned a randomly initialised model vector that has the same dimensionality as the input vectors. In each iteration, a randomly selected vector is matched with the closest model vector (winner). An adaptation of the model vector is performed by reducing the distance between the model vector and the feature vector. The neighbours of the winner are adapted as well, yet to a lesser degree than the model vector of the winning unit. Once the learning phase is completed, the feature vector of each music file is mapped to its best-matching unit on the map. The axes of the map have no specific meaning, the information is conveyed through the distances among the music files to each other.

One of the application scenarios of the resulting SOM is, in conjunction with different visualisations, as a *Music Map* giving an interactive and intuitive overview over a large audio collection. Furthermore, this allows to quickly create various playlists by drawing a path on the *Music Map* (see Figure 1). The system selects tracks that are located along the path, creating a sequence of songs following the "geographical" direction of the path. This results in a playlist re-

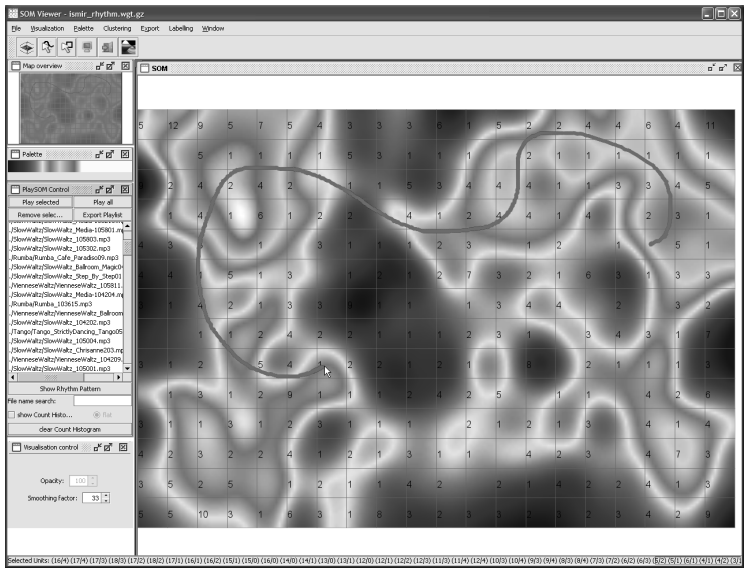


Fig. 1. PlaySOM showing a *Music Map* with a path and the corresponding playlist

stricted to a very specific musical style (which can, of course change over the progress of the playlist as the path moves on to some other region) but still containing a certain amount of variance. [4, 6]

3 Visualising Playlists

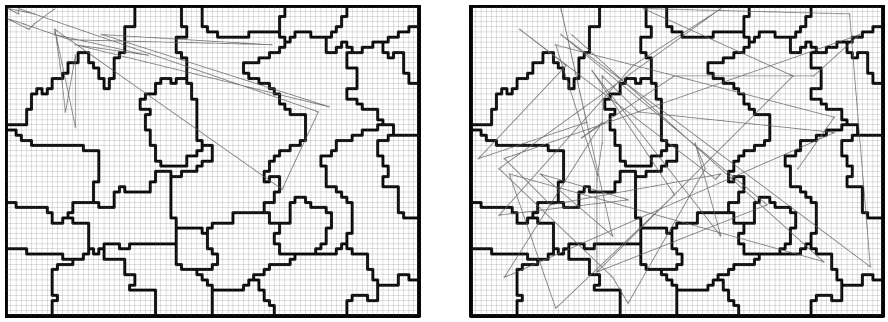
To visualise a playlist on the *Music Map*, for each song on the given playlist, the position on the map is located and linked with the positions of the adjacent songs oin the playlist. This creates a path based on a playlist. Such visualisations can be used as a template for new playlist (showing them as example or recommendation to the user), but they also reveal specific and descriptive information about the playlist. In conjunction with different visualisations of the *Music Map* the shape of the playlist and the regions it covers allows conclusions about the musical style and variance of the playlist.

3.1 Data Corpus

To analyse the visualisation of playlists, a data corpus has been created, based on playlists created by last.fm[1] users.

Last.fm users have the posibility to add songs they like to one of their personal playlist, which is stored together with their user profile. As users can view the profile of each other, it is also possible to view the playlists of a certain user.

[1] `http://last.fm`



(a) Homogenous and focussed playlist (b) Playlist without clear focus

Fig. 2. Two playlists, visualised on a *Music Map*

Furthermore, it is also possible to listen to playlists that contain at least 45 tracks from at least 15 different artists[2].

The data corpus was created by randomly selecting approximately 900 last.fm users who had created at least one playlist with more than 20 songs. The playlists where fetched and saved, together with 30 second snippets of the of the songs which are available from last.fm. 30 second preview snippets are also available at amazon.com[3], so songs that could not be retrieved from last.fm were fetched from this alternative source.

This resulted in a data corpus of 31772 audio files and 960 playlists. A *Music Map* with 80×60 units was created for the visualisation and analysis of the playlists. For all experiments in this paper, only playlists with a coverage of 100% where used – reducing the size of the courpus of playlists from 960 to 82 playlists.

Finally, the corpus was further extended by adding playlists created by professional users. The song titles broadcasted by a popular mainstream pop radio station[4] where logged over several days and again 30 second snippets where fetched from last.fm or amazon.com. The playlists retrieved from the radio station where split at the full hour to avoid the gap created by news broadcast.

3.2 Analysis

The visual analysis of the playlists showed very heterogenious results. Some playlists clearly focussed onto a specific region on the map. Figure 2(a) shows an example of a playlist with a "good" shape, that stays in an area of the map with infrequent outliers with a larger distance.

[2] At the time of investigation, this feature was reserved for users willing to pay a monthly fee.

[3] `http://www.amazon.com`

[4] Oe3, `http://oe3.orf.at`

Figure 2(b) provides as an excellent counter-example. The playlist covers almost the entire map, jumping from one corner to the complete opposite. It further shows no clear focus for any region on the map.

Unfortunately, playlists of the latter shape are the clear mayority. Only a handfull of playlist formed a continuous path. Further investigations of the playlists and their paths showed, that playlists that focussed on specific regions in most cases only contained tracks from one artist, sometimes even only from one album, which is uncommon for most playlists, thus raising the questions whether the playlists retrieved from last.fm can be considered as "real" playlists, or if they are used as some kind of "bookmark list" for favourite songs, but, on the other hand, many reached from one end of the map to the other.

The visual appearance of the radio playlists was, however, about the same as the previous investigated playlists from last.fm. The main difference was that none of the radio playlists was homogenous enough to really focus on one specific area.

4 User Study

To further investigate the quality of the playlists from last.fm, and to determine whether the visual shape of a playlist allows conclusions about the quality of a playlist, a small scale user study was launched. Two playlists where chosen from the pool of available last.fm-playlists based on their graphical representation on the map, one with a "good" visual shape and one with a "bad" shape (covering a large part of the map, c.f. Figure 2). Further also two playlists from the radio station where picked analogous to the playlists from last.fm. Finally, two playlists generated through the map (by drawing a path onto the map) where included in the questionaire.

Before given to the participants, all playlists where truncated to equal length (of 13 songs) and their names and therefore their sources where concealed. For each of the six different playlists, the participants where asked to

- rate how good a song fits to its preceding song, on a scale from 1 to 5, where 5 represents the best value,
- give an overall rating of the playlist on the same scale,
- select up to three songs from the playlist that should be removed to improve the quality, and
- name situations and/or locations where this playlist would fit in.

4.1 Participants

The questionaire was completed by five participants (three male, two female). A user study in this size must not be considered as a representative evaluation, it is more a proof of concept for the questioning.

The participants where between 20 and 30 years old. All of them have attended natural science studies at a university; three of them already received a master's degree, one is still a student and one droped out. Four of the participants are affiliated with a university or research institution.

Table 1. Average transition ratings by the participants. t is the average transition rating, o is the overall rating of the playlist.

Playlist	P-1 (m)		P-2 (m)		P-3 (f)		P-4 (m)		P-5 (f)		average	
Playlist	t	o	t	o	t	o	t	o	t	o	t	o
last.fm A	2.25	2	1.67	1	2.00	3	2.17	2	2.67	2	2.15	*2.0*
last.fm B	2.75	3	2.17	3	2.92	1	3.00	3	3.67	4	2.90	2.8
radio A	2.50	4	3.17	4	2.50	4	1.67	1	2.92	3	2.55	3.2
radio B	3.42	4	3.33	4	3.00	3	2.83	2	4.33	5	**3.38**	**3.6**
generated A	2.17	2	1.67	2	2.08	3	1.67	2	2.75	2	*2.07*	2.2
generated B	2.67	2	2.42	3	3.50	4	2.33	3	3.33	4	2.85	3.2

4.2 Results

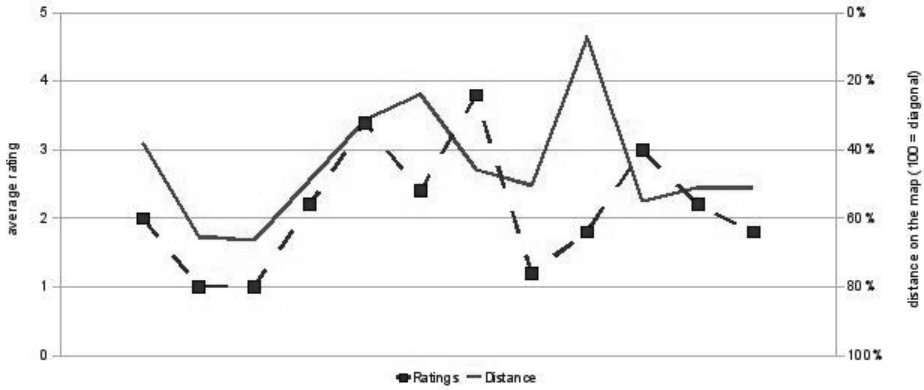
From the questionaires of the user study, summarised in Table 1, several interesting conclusions could be drawn:

1. Individual participants gave coherent answers. In most cases, the average rating of the transitions was about the same as the overall rating of the playlist.
2. The participants showed a clear preference for "professional" playlists taken from a radio station over all other playlists (c.f. Table 1).
3. Regarding the overall rating, generated playlists performed better than the playlists from last.fm. For the transition-rating, the playlists from last.fm where slightly in favour.
4. Regarding playlists from the same source, the one with the "better" visual shape also got higher ratings in all cases.
5. Whereas the individual rating of the transitions differ in most cases, the participants clearly agreed when naming the songs that would not fit into the playlist and thus should be removed.

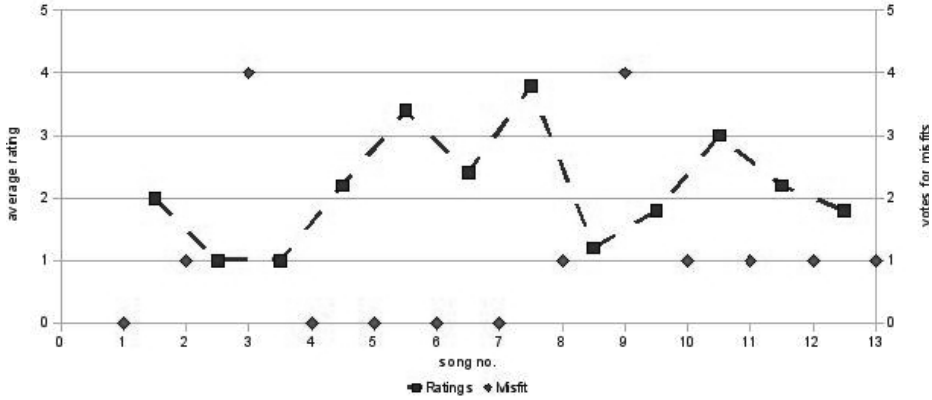
For all feature sets that have been investigated, no correlation between the users' rating and the distances in the feature space could be found. This observation can also be made for the distances on different maps trained with these feature sets.

This is illustrated in Figure 3 for a map based on Rhythm Patterns: in 3(a), the dotted line represents the average ratings by the participants (left Y-axis), while solid line represents the distances in the input space (right Y-axis, 100% is equivallent to the map's diagonal). The third song in the playlist has a notable longer distance to its neighbours in the list – these transitions also were rated rather low. The middle part of the playlist could be interpreted that closer distances got better ratings, but towards the end occurs a section where the closest distance in the playlist got very low ratings.

An interesting observation from Figure 3(b) ist the correlation between the ratings and the votes for misfitting tracks. In most cases, where the mayority of the participants found that a song does not fit into the playlist, also the transition to and from this song got notable lower ratings. Again, the playlist shown is exemplary, but the trend can be observed with all playlists.



(a) Partially correlation between the distance on the map and the users' ratings



(b) Transitions to and from outliers are rated low

Fig. 3. Graphical representation of one exemplary playlist *last.fm B*

5 Conclusion and Future Work

In this paper, the correlation between a playlist and its visual shape on a *Music Map* was investigated. It showed, that very homogenous playlists (by one artist or album) do stay very focussed in specific regions, but playlists with more diversity are distributed over large parts of the map. This is especially true for professional playlists. Future work on this part will be to investigate the behaviour of playlists on other maps based on a bigger, more heterogenious corpus.

The user study showed that the visual shape in the *Music Maps* used in this paper is not a sufficent quality measurement on its own, but it does allow conclusions regarding playlists from the same source. Furthermore it showed that also the feature vectors do not cover all aspects of playlist quality. It will be part of further investigations which additional aspects are required to sufficently describe the quality of playlists. Part of this work will comprise a refined user study with a more balanced and representative participants' list.

References

1. Jakob Frank. Enhancing music maps. In Frantisek Babic, Jan Paralic, and Andreas Rauber, editors, *Proceedings of the 8th International Student Workshop WDA 2008*, pages 53–59, Dedinky, Slovakia, June 26-29 2008.
2. Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995.
3. Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 34–41, London, UK, September 11-15 2005.
4. Robert Neumayer, Michael Dittenbach, and Andreas Rauber. PlaySOM and PocketSOMPlayer – alternative interfaces to large music collections. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 618–623, London, UK, September 11-15 2005.
5. Andreas Rauber, Elias Pampalk, and Dieter Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 71–80, Paris, France, October 13–17 2002.
6. Andreas Rauber, Elias Pampalk, and Dieter Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003.

Session 2

Digital preservation

Aspects of a Small Scale Long Term Archiving System

Stephan Strodl

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/~strodl
strodl@ifs.tuwien.ac.at

Abstract. Large heritage institutions have been addressing the demands posed by digital preservation needs for some time, but in contrast, small institutions and private users are less prepared to handle these challenges. Nevertheless, each has increasing quantities of data that exhibit considerable value. The Hoppla archiving system is designed to be used in environments where little professional know how or awareness of digital preservation issues are available. Two key aspects of the software implementation are presented in this paper. The data model provides a core module of the system and the preservation action workflow implements the logical preservation of the archiving system.

1 Introduction

Digital information is of crucial value to a range of institutions, from memory institutions of all sizes, via industry and SME (Small and Medium Enterprises) down to private home computers containing office documents, valuable memories, and family photographs. While professional memory institutions make dedicated expertise and resources available to care for their digital assets, SMEs lack both the expertise as well as the means to perform digital preservation activities to keep their assets available and usable for the future. Small institutions are starting to pay attention to the long term availability of their documents for business interest as well as for legal obligations.

In order to address the issue, we created a concept for a small scale archiving system. The system design and requirements for such a system are presented in [3]. The fully-automated preservation system was developed to be used in environments where little professional know how or awareness of digital preservation issues are available. The underlying principle of the archiving solution is finding a best effort solution with respect to the available technology, resources and skills of the users. We cannot assume a highly sophisticated computer environment; neither can we expect profound knowledge in digital preservation or archiving.

With the software Hoppla we are currently developing a prototype implementation of the system. It combines back-up and fully automated migration services for data collections in small institutions and personal settings. It combines

bit-stream preservation via the LOCKSS-principle [1] with logical preservation by automatically obtaining migration rules and tools. This allows outsourcing of required digital preservation expertise for the institutions. The system aims to provide the best available and most practical preservation solution base on expert advice via an automated web-service-based update of preservation plans.

In this paper we are presenting two aspects of the Hoppla software in detail, the data model representing a core concept of the whole system and the preservation action allowing the logical preservation of the collection. Furthermore the updating mechanism for outsourcing preservation expertise is presented. The novelty and complexity of the system raises a number of research questions, some of them addressed by planned future work are presented in last chapter of this paper.

The remainder of this paper is organised as follows: Section 2 provides pointers to related initiatives and gives an overview of work previously done in this area. Aspects of the software implementation including the data model and the preservation action are presented in Section 3. An outlook on future work is presented in Section 4.

2 Related Work

A number of research initiatives have emerged in the last decade in the field of digital preservation, primarily memory institutions focusing on professional environments. The raising awareness for small institutions and SOHOs increases demand of practical solutions for users with less experience [1]. Small institutions can benefit from projects for professional setting.

Existing open source digital repositories, such as Fedora Commons[2] and DSpace[3], are developed for large scale collections in professional archiving. These repositories provide a huge function range, but require considerable knowledge for configuration and usage. The overhead of function and configuration make these systems unsuitable for institutions with limited knowledge in data management. The innate support of these systems for logical preservation is limited. Considerable effort of development would be necessary to provide long term preservation functionality for a collection. The monolithic design of the systems makes adaption and customisation of the core system a difficult task.

Research on migration as a technical preservation strategy was done by the Council of Library and Information Resources (CLIR). They presented different kinds of risks for a migration project [2]. Migration requires the repeated conversion of a digital object into more stable or current file format. Migration is a modification of the data and always incurs the risk of losing essential characteristics of the object [2]. Therefore, a verification of completeness and correctness of the migration activity is required for a preservation system. Still the number

[1] `http://www.lockss.org`
[2] `http://www.fedora.info`
[3] `http://www.dspace.org`

of tools as well as the ease of applying migration makes it a very promising candidate for archiving in small institutions.

3 Software implementation

The aim of the prototype software implementation of Hoppla is to provide a solid framework for testing and evaluating different modules and techniques of an automated archiving system. During the development a high number of challenges occur that had to be solved indicating the complexity of the system. The current software system allowed us to test a first set of functionalities and to validate parts of our design, for example the first version of our data model design.

A key requirement for the software system was the strict modular design. It required a lot of design effort for the software components and their interfaces. The modular design allows the integration of existing modules or systems into the Hoppla system. Another aspect of modular design is the plug-in infrastructure for acquisition and storage media that ease the integration of other source and storage media.

In this paper two aspects of the software implementation are presented. At first, the data model that forms a core concept of the archiving software is discussed. It needs to support the flexibility and the modular design of the whole archiving system. The second section focuses on the preservation workflow describing the required steps to perform migrations.

3.1 Data model

The data model forms a critical part of the whole system as all other module are using the data structure. The great challenge of the data model for the archiving system is to provide a high degree of flexibility to support all future changes and extension of the system (for example plug-ins for new storage or acquisition media or the use of new format identification services). The different structures of metadata also require flexible data structure, they depend on the format and the used characterisation tool to extract the metadata. On the other hand the data model needs to provide a static data structure that access and retrieval functions can process, interpret and use the data. In order to meet the requirements specific data entries can be extended by predefined data structures. For example an extraction tool for PDF can add a data entry for the page count.

Moreover, the data model needs to support the life cycle of an object including versioning, migration and backups. For backups, the data model needs to manage the multiple storage locations of versions and migration.

Figure 1 shows the core data entities of Hoppla's data model and their relationships. The data model abstracts the concepts of Element, Version, Migration and Manifestation. The entities are described in detail below.

Project encapsulates one preservation effort, e.g. preserving all private documents and emails. A project is made up of several sources and storage media.

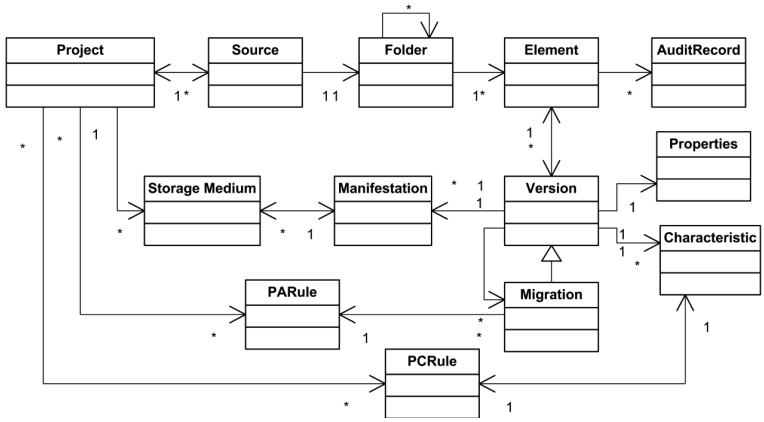


Fig. 1. Abstract Data Model

Source represents a source location of data for preservation, e.g. local folder or e-mail account. Where adequate the elements from the Source are mapped to a Folder/Element paradigm. It should ease the retrieval of elements and preserve the order of the elements on the local systems.

Folder as in a local filesystem a Folder can hold several subfolders or elements.

Element represents an 'intellectual' object (e.g. photos, text document) that can be represented by different physical representations (in data model versions and migrations).

Audit Record contains the documentation of all actions that take place in the archive.

Version is a specific unique version of an Element that was ingested into the Hoppla System.

Manifestation represents the physical manifestation of a specific version or migration on a storage media.

Storage Medium specifies storage devices for a Project. Hoppla differs between write once and re-write media.

Migration each Version of an Element can have different migration, e.g. Migration of a Version in Microsoft Doc format into txt and Pdf/A. Migration is implemented as an extended class of Version.

PARule specifies a preservation strategy. It includes the input objects (e.g. specific format), the PA tool and the parameter setting. More details about the PARule are presented in Section 3.2.

Properties describe the general metadata of a Version including format, size and source system properties.

Characteristics Metadata extracted from a Version or Migration using a PCRule.

PC-Rule specifies a characterisation service using a specific PCTool, for example using JHOVE to extract metadata form a PDF.

3.2 Preservation Action

Preservation actions in Hoppla are guided by migration rules. The rules specify migration strategies for specific objects in a collection. They are provided via the Web Update Service to the Client. These rules and the according preservation tools are frequently updated. In this sense, Hoppla acts similar in principle to Antivirus software. It regular updates local databases with rules and patterns to detect viruses, as well as with tools to automatically remove them. Similarly, Hoppla retrieves migration rules and tools with settings optimized by experts based on best-practice experience and detailed preservation planning following well-documented procedures. In order to reduce the risk of losing essential characteristics of the object by migration, validation rules are assigned to preservation rules. Where possible the results of the migration are checked for completeness and correctness through validation services. In the current version of the software only minimal version of the validation is implemented.

Preservation action in Hoppla needs to define the following abstractions.

Tool A tool is an executable application consisting of one or more files. A Tool can have one or more Configurations, for example ImageMagick convert.

Configuration A Configuration encapsulates execution parameters which determine the way the tool will be executed, for example parameters to set a specific resolution or colour depth.

Constraints A Configuration can have zero or more Constraints which define under which conditions a certain Configuration can be used to migrate an object. Examples for such constraints are usually filetypes, max/min object size, or the presence or lack of certain technical characteristics such as e.g. no transparent image layers.

Rule A Rule is a data set consisting of a tool-ID, a Configuration and Constraints. The Rules are transferred from the Web Service to the client defines a preservation strategy for preserve specific objects.

Migration Workflow

Preservation Management is responsible for the logical preservation of the collection. In terms of Hoppla, a great challenge for the implementation of the migration workflow is the multiple sources of potential errors and the required error tolerance of the system. The use of external tools always bears the risk of failures, especially migrations that include the modification of the data are fault-prone. Therefore the implementation needs a solid error handling.

The migration workflow consist of the following steps, which are described in the follow in detail,

1. Creation of Collection Profile
2. Request of Preservation Rule and Tools
3. Tool Preparation
4. Execution

1. Creation of the Collection Profile

The first step is to create a collection profile that describes the collection and its

characteristics. Hoppla supports different levels of detail of the collection profile. The profile can range from a list of object formats to detailed statistics on the number of objects per format including size and specific technical characteristics. For privacy issues, the user can select the level of detail sent to the web services. Detailed profiles allow more precise preservation rules for the collection, for example images with transparent layers require the use of special migration tools that support transparent layers.

2. Request of Preservation Rule and Tools
The client sends the collection profile containing information about the objects present locally to the Web Update Service. Based on the collection profile the Web Update Service recommends a list of preservation rules for the client. The first version of the selection of the preservation rule will be performed by pattern matching by using the predefined levels of preservation. In the first version of Hoppla the following preservation levels are available

- **Essential preservation** Obsolete formats that are at imminent risk of loss are migrated.
- **Recommended preservation** These rules are based on experiences and recommendation from preservation experts. They include migration of formats that are no longer in wide-spread use, outmoded (e.g. old versions of application) or the migration of proprietary format to open source formats.
- **Pro active preservation** Pro active preservation rules are based on best practice and providing additional multiple migration pathways for object formats to support different future usage scenarios, e.g. migration of Word objects to open document format and PDF (preserving the layout and the editability).

The result of the second step is a list of recommended preservation rules that is send to the Client.

3. Tool Preparation
Preservation tools that are required to perform the preservation rule are downloaded via the Web Update Service. Hoppla supports three different kinds of migration tools.

- **Portable tools**
  They do not need any installation on the target system (e.g. java classes or statically linked executables). The tools are downloaded from the server and stored on the client side. All portable tools from the Hoppla Web Service provide an unique interface for the client side to set the parameter and execute the tool.
- **Installed tools**
  In order to use a wider range of preservation tools, Hoppla implements a small discovery service that searches for installed software on the client system. Command line execution is the preferred way to search, but a registry

search for MS Windows is also implemented. The list of tools found by Hoppla is provided to the web service. The Web Update Server can provide preservation rules that use installed tools. For privacy reasons the user can edit the list of tools provided to the server. For each installed tool a small wrapper needs to be downloaded from the web server. The wrapper implements a unique interface that allows the client to use the tools.

– **Web Services**
They can be used if licensing or platform incompatibilities hinder installing the software on the client. Privacy, security and performance issues have to be considered. Web Services are currently only considered in design perspective, but are not implemented at this stage.

In order to manage the dynamic tool integration Hoppla implements a Tool Manager. The tool manager supports the identification and initialisation of downloaded migration tools. Therefore the folder where the downloaded tools are stored on the client side is scanned at the application start. The identified tools are instantiated and can be used by the Hoppla on the client side.

4. Execution
The first step is to determine the objects to migrate in the collection. The constraints defined in the preservation rule are used to determine the objects. The metadata repository of Hoppla is queried with the constraints to identify all the objects that are covered by the preservation rule. The migration itself is performed in a sandbox environment. When available the migration results are checked for correctness and completeness through validation services. Correctly migrated objects are ingested into the collection and stored on the target media.

4 Outlook

Hoppla presents a new concept for archiving system providing highly automated workflows and outsourcing of preservation expertise. The first software prototype implementation allows a first feasibility evaluation and validation of the system design. It further supports the refinement of theoretic concepts and designs. This paper presents the first versions of the data model and the preservation action workflow.

As the archiving system is complex and covers a wide range of functionality, the current version has limitations and restrictions. Further development effort will extend and refine the functionality of the software. The software implementation should build a solid basis to test and evaluate research aspects and experimental techniques of the system.

Research work will consider detailed analysis of the qualitative trade-off between automated and expert-guided digital preservation solutions. A special focus will be on preservation planning in the web update service based on different requirements from the client (e.g. user profile, system profile). Other aspects related to preservation are techniques to automate the validation of migration results.

A further research aspect is complex objects, in this context the data model will be analysed how far it can be extended to support complex objects. The different kinds of relationship need to be further investigated and their effects on the life cycle actions.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project (contract 033789) and by the Austrian Research Promotion Agency (FFG) through the Research Studio Digital Memory Engineering.

References

1. Bradley, K. Digital preservation: the need for an open source digital archival and preservation system for small to medium sized collections. `http://portal.unesco.org/ci/en/files/28067/12323631793BradleyPaper.pdf/BradleyPaper.pdf`, 2008.
2. Lawrence, G. W., Kehoe, W. R., Reiger, O. Y., Walters, W. H., and Anne, K. R. *Risk Management of Digital Information: A File Format Investigation*. Council on Library and Information Resources, 2002.
3. Strodl, S., Motlik, F., Stadler, K., and Rauber, A. Personal & SOHO archiving. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL'08)* (Pittsburgh PA, USA, 2008), ACM, pp. 115–123.

Challenges for the Evaluation of Emulation

Mark Guttenbrunner

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/dp
guttenbrunner@ifs.tuwien.ac.at

Abstract. Emulation is one of the main strategies for preserving digital objects. With different emulation environments to choose from when preserving a collection of digital objects it is necessary to choose the emulator that is able to render the significant properties of an object best. But also on the development side it is necessary to run automated tests for new versions of emulators.
In this work we present some of the main challenges for automated testing of emulators with specific digital objects. To compare different original and emulated environments it is necessary to document the environment used to render an object. Changes in this 'view-path' comprising all secondary digital objects needed to perform the emulation can lead to changes in behavior or appearance. For interactive and dynamic objects it is necessary to determine the external events influencing an object, e.g. user input and changes in synchronized timing can lead to different results when executing a digital object.
Significant properties of descriptive forms, i.e. the physical manifestations of an object, are compared when using migration as a strategy by comparing the original and the migrated form. With emulation the descriptive form of the object remains unchanged, so properties of a rendered version of the object have to be compared to determine if the preservation was successful. Emulation environments have to be extended to allow for the extraction of significant properties of digital objects. These properties can vary over different points in the rendering process.

1 Introduction

Research in preservation planning and evaluation of tools and objects has been made mainly for migration as a digital preservation strategy. But migration is not always a suitable strategy. For dynamic and especially interactive digital objects emulation is an important strategy as well. It is well known how to extract and compare significant properties for migrated objects, but with emulation the original object is unchanged. Instead of comparing an original and a migrated version of a digital object the comparison of a rendered version of the object in its original and in an emulated environment is necessary to determine if the significant properties of the object stay intact.

Similar to the migration of digital objects the goal of the evaluation of emulation environments is to perform repeatable experiments that allow us to take

an informed and accountable decision on the best emulation environment for a certain digital object. To achieve this we have to extract significant properties from the rendering process. Automatic comparison of significant properties extracted from different emulation environments should be made possible. These steps would allow us to do preservation planning for emulation environments and automate parts of the process of testing emulators.

This article shows the challenges of evaluating emulation environments. It is structured as follows. First an overview of related work is given in Section 2. Then the importance of rendering properties for different types of digital objects for which emulation could be a suitable preservation strategy is shown. Next we discuss the necessity of documenting the environment and automating external events. We take a look at the rendering process and the extraction of significant properties from the emulation environment. Finally in Section 8 we present the conclusions and discuss what future work has to be done.

2 Related Work

Migration ([10]) and Emulation ([7], [12]) are listed in the UNESCO guidelines for the preservation of digital heritage [13] as the main strategies for digital preservation. Emulation refers to the capability of a device or software to replicate the behavior of a different device or software. It is possible to use hardware to emulate hardware, software to emulate software or software to emulate hardware. The challenges presented in this article mainly use *Emulator* as defined in [8] for a program that virtually recreates a different system than the one it is running on, but most of the presented work applies to Emulation in a broader context as well.

Previous research has been done on methods for evaluating the effects of migration on documents. A preservation planning workflow is described in [9] and allows for repeatable evaluation of preservation alternatives. An implementation of this workflow has been done in the preservation planning tool *Plato* ([1]), utilizing automatic characterization of migrated objects with tools like *Droid* ([3]) to identify files. The significant properties of migrated objects can be compared automatically using the *eXtensible Characterisation Language* (XCL) ([2]) to measure the effects of migration on the object. While the preservation planning tool can be used to compare emulation environments as shown in a case study in [5], the comparison has to be done manually. Significant properties of software as one category of dynamic objects are listed in [6]. In [11] the information contained within a file is distinguished from the rendering of this information. To compare emulation environments we have to compare information about the rendered object as the object itself is unchanged in different environments.

3 Rendering Properties for Object Types

In [4] the following types of interactive digital objects are described as candidates for using emulation as a digital preservation strategy: application software,

dynamic documents, interactive art and video games. For these object types emulation is an obvious choice to preserve the interaction properties. But all other non-interactive digital objects can be candidates for preservation through emulation as well. If objects have to be kept in the original format (e.g. for legal reasons) even for static documents emulation might be the strategy of choice.

Especially for dynamic objects properties of the rendering process that are not encoded in the descriptive form, i.e. the physical manifestation can be relevant as well (e.g. frame rate, CPU-cycles in a certain amount of time, number of disk operations, reaction time between user input and resulting change of the rendered object). These significant properties have to be determined for an object or a class of objects that have to be preserved.

4 Documentation of Environment

One of the steps in the preservation planning workflow defined in [9] is to describe the conditions under which a migration of objects was performed. For emulation this is a crucial step as it is necessary to define the environment in which the object is executed in. Every setting of the rendering system can influence the behavior and appearance of the object. Besides the settings for hardware, operating system and the digital object itself, other digital objects influencing the objects rendering process (e.g. additional software on a system influencing the speed, operating system plug-ins effecting the appearance) have to be considered as well.

For every digital object a view-path of necessary secondary objects can be constructed. Secondary digital objects are software that is needed for rendering the object that has to be preserved. As an example an operating system and a viewer application might be the minimal view-path to render an image. As the same image can be rendered using different viewer applications not necessarily running on the same operating system, results of the rendering process can be different (Figure 1). Depending on the object environment settings can influence behavior and appearance. While video games usually have an appearance that is independent from settings, the look and feel of application software could be changed in the operating system configuration.

To minimize the side effects of changes in environment on rendering, the view-path with all relevant settings should be well documented. This lays a foundation for evaluating emulators. By keeping the view-path constant we can make sure that differences in rendering are caused by differences in emulation environments and not by changes in the view-path.

5 Automating External Events

Depending on the object type different external events influence the behavior or the appearance of an object. Changes in these events between the original environment and an emulated environment can change the behavior of an object. Three major influences on dynamic objects are listed below.

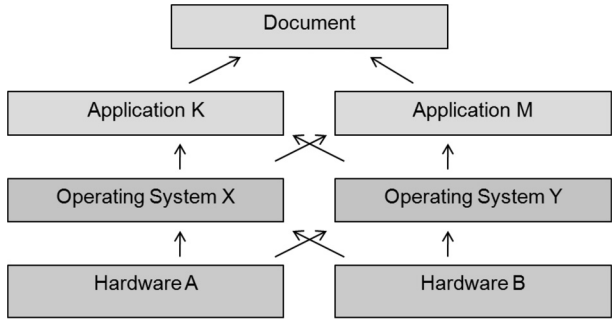


Fig. 1. Different view-paths for rendering a digital object.

5.1 User Input

For interactive objects changes in user input influence the behavior of an object. To compare the behavior of an object in the original and the emulated environment for testing purposes we therefore have to unify not only the inputs that were triggered by a user but also the precise moment at which those inputs occurred. Depending on the object the timing might be more or less crucial. A mouse click that is 1 second to late might not make a difference for a database application, while it completely might change the events in a video game or a piece of interactive art.

As it is impossible to manually apply the same interaction twice at the precise same moment during the emulation process, it is necessary to automate the application of user input. This is currently either not supported by emulators or supported in proprietary format. Input scripts recorded on one environment (original or emulated) cannot be applied to other emulation environments.

5.2 Synchronized Timing

Behavior of objects can be bound to time variables. These are usually hardware specific registers that can be read using software. Especially for early video games the timing of graphic objects is often bound to the location of the raster beam on the screen. Differences in emulation of the timing and the original timing can lead to unwanted differences in the displayed image. But not only the raster beam, also processor interrupts or hardware system clocks are potential causes of changes in the rendering process for time critical code.

To avoid side effects due to differences in timing it is necessary to synchronize timing to the start of the emulation process, e.g. a hardware clock has to be set to the same value in different emulation environments to synchronize the events instead of using the hardware clock of the host system.

5.3 Randomization

Especially video games and interactive art objects use random events. The occurrence of random events is usually tied to a random number generation. The

algorithm that generates the number has to be initialized by setting a base number that is set using an indeterminate hardware number (e.g. hardware clock, raster beam, timed user input).

By identifying the source for the random number based algorithm and keeping the generation of the number constant over different emulation environments it is possible to eliminate the randomness of the algorithm.

6 Rendering

Significant properties of an original and a migrated version of the descriptive form of an object are usually compared when using migration as a digital preservation strategy. If the migrated version has the same significant properties as the original version, then the tool used to migrate the object is considered a successful preservation action. For emulated environments as preservation action tools the situation is different. The descriptive form of the object is unchanged for the original environment and different emulated environments. Rendered version of the object have to be compared instead of descriptive versions.

In an environment used to render the object usually various different forms of the object exist. It could be stored and rendered in a location in system memory after being loaded by a viewer application, stored in specific memory of the outputting hardware (e.g. video card) or output on the output interface of the system (e.g. display device). To compare an object in different environments the same rendered version of the object has to be used and extracted from the environment.

7 Extraction of Significant Properties

Rendering of a dynamic object is a continuous process. The significant properties of an object can change likewise during the rendering process in different states. Depending on the object some of these states might be significant while others might not be. A static document will have only one significant state once it is loaded and displayed while loading a series of web-pages creates various significant states when the different web-pages are displayed. In the case of a video game every rendered frame on the screen can be significant to preserve the changing image. Extracting properties therefore can also be done at one moment of emulation, at various states during the execution of an object or as a continuous stream.

With migration as a strategy it is possible to create tools that examine the migrated object and extract significant properties. For emulation the properties have to be extracted from the rendering process. As not all information about the rendering process is visible on the rendered object it is necessary that the environment supports the extraction of defined properties.

8 Conclusions and Future Work

In this paper we discussed the challenges and requirements for the evaluation of emulation environments in the context of digital preservation and for selection of the best emulation tool for a set of digital objects.

Depending on the object type not only the properties of an object stored in the descriptive form, but also properties of the rendering process can be relevant. By documenting the reference environment in which the object is originally executed along with all secondary objects potentially influencing its rendering we can minimize side effects on appearance and behavior due to changes in the environment. The significant properties of the object have to be defined and random elements have to be made deterministic to create constant behavior and appearance. Decisions have to be made on what level to compare the rendered object and in what regularity (once, at specific times or continuously).

Future work should be done on the development of guidelines for the support of automated user interaction in emulators. Significant properties that can be extracted from the emulation environment have to be defined and mapped to a characterization language like XCL. Emulators should include the possibility to extract information in the defined format as well as support for the extraction of rendered forms of the object at specified points of emulation.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

1. Becker, C., Kulovits, H., Rauber, A., and Hofman, H. Plato: a service-oriented decision support system for preservation planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'08)* (June 2008), ACM.
2. Becker, C., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. Systematic characterisation of objects in digital preservation: The extensible characterisation languages. *Journal of Universal Computer Science 14*, 18 (2008), 2936–2952. `http://www.jucs.org/jucs_14_18/systematic_characterisation_of_objects`.
3. Brown, A. Automatic format identification using pronom and droid. *Digital Preservation Technical Paper 1* (2008). `http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf`.
4. Guttenbrunner, M. Preserving interactive content: Strategies, significant properties and automatic testing. In *Workshop on Data Analysis (WDA'2008)* (Dedinky, Slovakia, June 2008).
5. Guttenbrunner, M., Becker, C., Rauber, A., and Kehrberg, C. Evaluating strategies for the preservation of console video games. In *Proceedings of the Fifth international Conference on Preservation of Digital Objects (iPRES 2008)* (London, UK, September 2008), pp. 115–121.

6. Matthews, B., McIlwrath, B., Giaretta, D., and Conway, E. The significant properties of software: A study. JISC Study, 2008. http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf.
7. Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources, January 1999. http://www.clir.org/pubs/reports/rothenberg/contents.html.
8. Slats, J. Emulation: Context and current status. Tech. Rep., 2003. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf.
9. Strodl, S., Becker, C., Neumayer, R., and Rauber, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)* (June 2007), pp. 29–38.
10. Testbed, D. P. Migration: Context and current status. White paper, National Archives and Ministry of the Interior and Kingdom Relations, 2001.
11. Thaller, M. Interaction testing benchmark deliverable PC/2 - D6. *Internal Deliverable, EU Project Planets* (2008).
12. van der Hoeven, J., Lohman, B., and Verdegem, R. Emulation for digital preservation in practice: The results. *International Journal of Digital Curation 2*, 2 (2007), 123–132.
13. Webb, C. *Guidelines for the Preservation of the Digital Heritage.* Information Society Division United Nations Educational, Scientific and Cultural Organization (UNESCO) – National Library of Australia, 2005. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf.

Digital Preservation Time Capsule: A Showcase for Digital Preservation

Natascha Surnic, Andreas Rauber

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/dp
{surnic,rauber}@ifs.tuwien.ac.at

Abstract. The importance of Digital Preservation is increasing steadily. The amount of digital data is growing fast, and maintaining the long term availability and accessibility of these data is turning into an increasing challenge. Creating an operational archival system requires to solve significant challenges in terms of analyzing and monitoring massive volumes of data while applying preservation actions. Still, simply demonstrating and grasping the issues surrounding Digital Preservation is non-trivial. It is very difficult to make digital preservation tangible, which is the main goal of the Digital Preservation Time Capsule. This paper presents the concepts of the Planets Digital Preservation Time Capsule, an appealing showcase demonstrating a range of activities in the context of Digital Preservation.

1 Introduction

The preservation of digital material has become very important. First of all the amount of digital data is growing fast and the material needs to be preserved. There are different ways to preserve digital material. It is possible to emulate the environment or to migrate data, for example. Another problem is that an appropriate storage media needs to be defined. To sum it up, we need to preserve the data in a long-term run. While numerous approaches and preservation actions have been devised, metadata schemas have been crafted, and storage solutions are on offer, really understanding the complexities of digital preservation, as well as the massive amounts of information required to be able to interpret digital objects, is a hard challenge. The Digital Preservation Time Capsule sets out to make these aspects easier to understand and to put them into context. The goal is to show the dependency of data objects required to turn them into information objects at all levels, starting from the basic primary objects via necessary repreentation information, software required to interpret the data, information rewuired to interprete the standards and other required documenation, the compilers for the viewer software and migration programs used, down to the actual operating system, and - eventually - the actual hardware architectures.

This article is structured as follows. An overview of related work is provided in Section 2. The objects to be placed into the time capsule are described in detail in Section 3. In Section 4 we discuss the expected results and give an outlook on future work to be done.

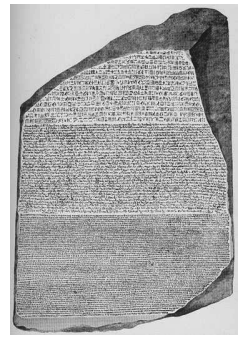


Fig. 1. Rosetta Stone [7]

2 Related Work

Digital Preservation has some very appealing and tangible showcases in the past event, if they may have been devised for different purposes. Examples are the Rosetta Stone or the Voyager Golden Record, which influenced the Digital Preservation Time Capsule.

2.1 Rosetta Stone

The Rosetta Stone ([7]), depicted in Fig. 1, is an Egyptian artifact which consists of Ptolemaic era steel with carved in text. The text is a single passage, a decree from Prolemy V, who described various taxes and instructions to erect statues in temples. The text is written in three different languages. Two are in Egyptian language, hieroglyphic and Demotic and one in classical Greek. The Rosetta Stone was created 196 BC and discovered by the French in 1799. Translations of the stone helped to decipher hieroglyphic writings. In OAIS [5] terminology, it provided vital descriptive information, albeit embedded, by offering several different representations of the same information. At its highest point the Rosetta Stone is about 114 centimeters high, approximately 72 centimeters wide and almost 28 centimeters thick. The approximated weight is about 760 kilograms. The Rosetta Stone is on public display at the British Museum since the year 1802.

2.2 Voyager Golden Record

The Voyager Golden Record [6], shown in Fig. 2, is a phonograph record, which was included in two Voyager spacecrafts. It was launched in the year 1977 and contains different sounds and images, which should provide an overview of the diversity of life and culture on Earth. It was developed for any intelligent extraterrestrial life form or far future humans that may find the record. The Voyager spacecrafts are not heading toward a specific destination but there are calculations where the spacecraft will be. The record is more seen as a Time Capsule,

Fig. 2. Voyager Golden Record [6]

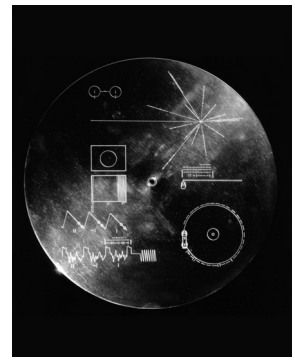


Fig. 3. Voyager Golden Record [6]

than a communication media with any other species because if they ever will be found, it would only be in far future. The content of the record were selected by a committee of the NASA. About 115 images, which are encoded in analogue form, and a huge variety of different sounds are stored on the record. Examples for the sound files are, different animal sounds, the sounds of wind and thunder. Furthermore there are spoken greetings in 55 different languages on the record. Also music is represented on the record. In a selection of 90 minutes, a variety of music from different cultures represents the sounds of the Earth. The Voyager Golden Record stores an hour long recording of the brainwaves of Ann Druyan. The record is a construct of gold-plated copper with an ultra-pure sample of the isotope uranium-238, which is electroplated on the cover of the record. This isotope has a half-life of 4,51 billion years and was chosen because if any civilization will encounter the record, it will be possible for them to determine the age of the record from the remaining uranium. The sentence "To the makers of music - all worlds, all times" is handwritten on the record itself. Voyager 1 was launched in 1977 and has already passed the orbit of Pluto in the year 1990, left the solar system in 1994 and is now in empty space. Similar to the spirit of the Digital Preservation Time Capsule, the Voyager Golden Record contains selected information items represented in such a way as to be presumably intelligible by other / future species.

The cover of the disc, cf. Fig. 3, provides an appealing example of trying to achieve communication with unknown consumers of the information. Instructions on how to "read" read the disc are displayed graphically, with the rotation speed of the disc, for example, being provided in time units of 0,70 billionths of a second, the time period associated with a fundamental transition of the hydrogen atom. The originating location is depicted as the relative position of our solar system with respect to 14 pulsars[1].

[1] http://voyager.jpl.nasa.gov/spacecraft/goldenrec1.html

2.3 LongNow foundation

The LongNow foundation [4] promotes awareness of the importance of long-term thinking and helps to understand the actual problem, its dimensions and to communicate these things beyond the core groups of experts, who are working on problem solving strategies. It created a series of projects [1] to demonstrate various aspects, such as the 10,000 Year Clock Project, the Long Server, or the Long Bets website.

3 Creating a Digital Preservation Time Capsule

In this section an overview of the activities of the Digital Preservation Time Capsule is given.

3.1 Selecting Primary Objects

A set of core objects, the Source Objects, will be selected. The selection of these objects will be small and should explain the core ideas within Planets. The Source Objects are:

Photograph in JPEG format which represents a popular format for images and is used by a vast majority of users of digital cameras. Photographs have been fascinating people since ages, that is the reason why we decided to use a photograph as a Source Object for the Time Capsule.

'Hello World' program in Java A simple and a few code lines similar to a 'Hello World' program in Java will represent the most elementary executable program in source code form.

Short movie on Digital Preservation showing the main aims of Digital Preservation in .MOV format which should act as a short training video for training events and deliver the key message of Digital Preservation to the public. Here are conformities in style with the Voyager Golden Record audio recordings.

Planets Project Homepage in HTML format acts as a representative of more complex object formats which do not only consist of a single file. The homepage focuses on simple, standard HTML-functionality, using formats, which are already within the scope of formats to be described.

Planets Brochure [3] in PDF format acts as a representative of one of the most wide-spread document formats. Furthermore it illustrates the key messages and goals of the Planets Projects.

3.2 Deriving Secondary Objects

The primary objects must be described by their technical and intellectual characteristics using a standard metadata scheme, like PREMIS [2].To guarantee a long term preservation, these objects must further be converted into a range of formats, which are more suitable for long-term preservation. For example, PDF

documents may be converted to PDF/A, or to an image format such as TIFF or JPEG2000. The Quicktime video may be converted to MPEG-4. Both, the tools used to perform the migration and the resulting objects need to be described using the metadata representation scheme, like the Primary Objects mentioned before. Furthermore, all file formats need to be documented by including the appropriate standards.

Viewer for the primary and secondary objects must be defined. The source and the documentation of the viewer must be documented too. In this case a simple PDF viewer, e.g. XPDF, a simple image viewer, e.g. GIMP, a simple video viewer, e.g. XINE, a minimalistic HTML browser, a minimal JAVA virtual machine and other compilers, all the way to a minimal operating system are needed. Further objects may be included. The list above represents only an excerpt of all the objects required to fully document the hierarchy of dependencies for any single digital file. After defining all the appropriate objects, these also need to be documented and described like the Primary Objects.

3.3 Storage

All the objects need to be stored. We choose a wide range of different storage media, like CDs, DVDs, USB flash storage, and HDDs. Conventional "analog" storage such as print-outs to paper, as well was hybrid storage on microfilm will be considered as well. In order to be useable, the storage encodings, the storage media and the reading technology need to be described using a standard metadata representation scheme as well. Descriptive information of the various data carriers, which are included in the Time Capsule, are given in this section. The data carriers should all carry, as far as possible, identical information. Exceptions are obviously planned for the print-to-paper and the microfilming. For some parts it is easier to represent the information in a digital way, instead of printing out a lot of pages to represent them. The achievement is, to include a rather large volume of representation information. Furthermore there will be more exceptions if we decide to include old data carriers, which have not been designed to carry such a large volume of data. To sum it up, the selected storage technologies encompass most types of storage principles, demonstrating storage capacity, size and the evolution of data storage during the years. The data carrier, the drive technology and the file systems must be described as well. This means, that the patent, the ISO Standard, and simple technical descriptions, as well as Wikipedia articles need to be included, adding to the list of secondary objects identified above. For example, the carrier of floppy discs, CDs, DVDs, Blu-Ray discs, HDDs, flash drives, selected tape type storage media and the drive technology description is needed. Furthermore, descriptions of the storage data carrier and their drive technology descriptions, microfilm description and the resolution settings for writing to microfilm need to be documented. An OLPC (One Laptop per Child) laptop is planned to be added to the Time Capsule, to represent a complete reading device. The laptop allows that all the objects can be rendered. The OLPC laptop serves as a host device for many other storage

technologies, especially, CD/DVD, external HDD, flash disk, as well as other devices, if they are connectible via USB. This is useful for exhibitions and for demo purposes of the Digital Preservation Time Capsule. All the objects, concerning storage, also need to be described and documented as the Primary objects.

3.4 Describing Metadata with PREMIS

All digital objects which are deposited in the Time Capsule will be described using a standardize meta-data container. At this moment, the plan is, to use PREMIS metadata, wrapped in a METS container.

3.5 Time Capsule as a showcase for Digital Preservation

The resulting material will provide a tangible showcase to the general public, offering itself for public exhibits at partner institutions, other libraries and archives, as well at museums of science and technology. The big achievement of the Planets Digital Preservation Time Capsule is to make Digital Preservation tangible. To achieve this goal, the Time Capsule will contain all the digital objects mentioned above, in multiple storage formats and encodings. The information will be stored on a wide range of different data carriers.

Objects contained in the Time Capsule The following list gives an overview of the planned physical items of the Time Capsule:

- several punch cards
- several floppy discs
- 1 or more archival grade CD
- 1 archival grade DVD
- 1 high-quality USB flash memory stick
- 1 external HDD drive
- 1 roll of microfilm
- paper print-outs
- 1 external CD/DVD/Blu-Ray reading device with USB connectivity
- 1 OLPC laptop

Deposit Locations for the Time Capsule The Time Capsule will be deposited in a prominent location to stress the key messages. Additional copies of the Planets Time Capsule will be deposit at partner institutions, memory institutions in general, libraries, archives and special museums, as long running test sets. The quality of storage carrier functionality and the quality of data reconstruction will be tested. The opening of the Time Capsules may happen in specific instances, like every 10, 20 etc. years. The primary deposit location of the Time Capsule should be an appealing, somehow exotic location, to offer people an emotional connotation with a longterm archive. There are two different types of deposit: On one hand, deposits that lock away a complete copy of the time capsule for a specific number of years; on the other hand a display deposit, which allows to show the concept of digital preservation at exhibitions with an open copy of the Time Capsule, to make Digital Preservation tangible.

4 Outlook

With different kinds of Media Activities, the Time Capsule should raise the awareness for Digital Preservation activities and challenges to a vast audience. Several activities, like press conferences and focusing on the goals and challenges in Digital Preservation, should motivate the general interest in Digital Preservation. Furthermore, exhibitions, at Planets partner institutions and/or science museums, can use additional copies of the Time Capsule for training purposes.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

1. Stewart Brand. *The Clock of the Long Now*. Basic Books, 2005.
2. PREMIS Editorial Committee. Premis data dictionary for preservation metadata (version 2.0). Technical report, PREMIS Editorial Committe, 2008.
3. The Planets Consortium. Preservation and long-term acess via networked services, June 2006. `http://www.planets-project.eu/docs/comms/Planets_Project_Brochure.pdf`.
4. The Long Now Foundation. The long now foundation, Aug 2009. `http://www.longnow.org/`.
5. ISO. Open archival information system - reference model (ISO 14721:2003). Technical report, International Standards Organization, 2003.
6. California Institute of Technology Jet Propulsion Laboratory. The Voyager Golden Record, Aug 2009. `http://voyager.jpl.nasa.gov/spacecraft/goldenrec.html`.
7. British Museum. The Rosetta Stone, Aug 2009. `http://www.britishmuseum.org/explore/highlights/highlight_objects/aes/t/the_rosetta_stone.aspx`.

Recommender Systems in Preservation Planning

Hannes Kulovits

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/~kulovits
kulovits@ifs.tuwien.ac.at

Abstract. Due to fast technological changes, digital objects face obsolescence quickly. This demands action from the person in charge to mitigate arising risks and ensure accessibility of the collection over the long term. To tackle this challenge different strategies such as migration and emulation have been proposed; however, the decision which one to adopt is complex and requires detailed knowledge and experience in digital preservation. The process of evaluating strategies against well-defined requirements is called preservation planning and shall support decision makers to opt for the right strategy. The result of this activity is a preservation plan which contains the decisions taken including the complete evidence base.
Even though tool support already exists, the creation of a preservation plan still is a complex and time-consuming endeavour due to demanding tasks such as finding potential preservation actions and eliciting the institutional requirements. This paper proposes conceptual enhancements in the planning process to integrate recommender systems in order to reduce the effort needed and enable preservation planning for users with less experience. Furthermore, the basis for the information filtering in the different areas is identified.

1 Introduction

The creation of a preservation plan encompasses several essential activities which must be accomplished in order to arrive at a well-informed, consistent, comparable, and accountable recommendation for a preservation solution. Preservation planning is the process of evaluating potential solutions against specific requirements and building a plan for preserving a given set of objects. To date this is mostly done manually and in a rather unstructured way with little or no tool support. In the course of Planets a preservation planning methodology and a software tool (Plato) implementing this methodology are being developed to address these issues. The methodology supports the evaluation of preservation strategies and the production of well-documented, accountable recommendations on which strategy to follow.

The current version of Plato supports the workflow and integrates services for content identification and preservation action. The software itself is built upon the Planets Interoperability Framework that guarantees loose coupling of services and registries through standardised interfaces.

Although the Planets preservation planning methodology is already well supported by Plato some steps are still very complex and require certain knowledge from the planner. Tasks such as discovering potential preservation actions to consider for evaluation, and selecting representative sample objects to apply them on are very challenging.

This paper proposes integration points for recommender systems in Plato to further support the users in their preservation planning endeavours. The integration of recommender systems in Plato strives for four major goals:

- Enable preservation planning for inexperienced users. Building the objective tree which lays at the heart of the preservation plan demands considerable knowledge regarding digital preservation from the planner. The requirements must be measurable, general enough to not focus on a particular preservation strategy and specific enough to reflect the institution.
- Reduce complexity of certain workflow steps. The selection of preservation alternatives for instance can be quite complex as numerous candidates might be available. Not only tools wrapped as web services must be considered as alternatives but also migration paths, emulation view-paths and tools not yet wrapped as web service. Furthermore 'do nothing' might also be considered as an alternative.
- Improve preservation action recommendation. As the quality of the recommended preservation action highly depends on the defined requirements to which the candidate solutions are evaluated against, particular attention should be paid to the process of requirements elicitation. In this stage a recommender component can advocate requirements that especially focus on the collection in question and are likely to be tautological regardless the institution.
- Reduce time and effort for planning activity. There is possibly a plethora of potential preservation actions available from which the planner has to choose for evaluation. The application of a recommender in this stage can support the planner by, for instance, filtering out non-applicable ones and selecting the top-N best performing services considering the institution's policy and requirements.

The remainder of this work is structured as follows. The next section outlines related work in the area of recommender systems. Section 3 gives an overview of the preservation planning workflow. Section 4 describes the integration of recommender systems in the planning workflow. Section 5 draws conclusions and points out directions for future work.

2 Related Work

In systems where the number of choices can be enormous, recommender systems [14] assist users in identifying a subset of items from a typically larger set of possibilities they might be interested in. The main goal of recommenders is to reduce the complexity for individuals trying to find their way through

large amount of information and suggesting those pieces that were supposedly the most relevant ones. Furthermore, recommender systems take personalisation into account which is targeted on the user, as the need for information differs for each user. A computer scientist for instance, is, probably more interested in software engineering books than books on marketing and sales - contemplated from a professional point of view.

A very common way to obtain recommendations is by word-of-mouth or by reading reviews about items one is interested in. Systems like amazon.com suggest products on that basis, depending on user profiles and recensions provided by known users.

In principle, recommender systems are seen from four main dimensions [12]:

1. How the system is modelled, i.e. how the recommendations are made.
2. How a recommender system is targeted, i.e. the level at which information is tailored.
3. How a recommender system is built.
4. How a recommender system is maintained (online vs. offline)

Evolving from information retrieval, recommender systems have traditionally been studied from the aspect of how the system is modelled. A classification that is commonly accepted distinguishes between collaborative filtering and content-based filtering [7] where the latter is deeply rooted in information retrieval. Content-based filtering is also known as cognitive filtering [10] and calculates similarities between a number of items a user appreciates, and the products that are not yet known to the user.

Recommender systems also often connect groups of users with similar preferences or interests to take advantage of the group's experiences. This somewhat circumvents today's recommender systems from typical information retrieval endeavours which primarily focus on the information-seeking goal of a certain individual. In contrast to content-based filtering, collaborative filtering calculates similarities between the users based on the available user profile. Collaborative filtering is also called social filtering [13].

A recommender system can also follow a combination of two or more different approaches (hybrid recommender) under a single framework in order to leverage the advantages of the individual one and address the disadvantages of each. Burke et.al. [5] surveys numerous ways of combining different approaches whereas the most popular amongst them is a composite of collaboration and content. One of the first hybrid recommenders is Fab [3] with the aim of proposing web sites to its users. Amazon for instance applies a hybrid recommender which combines content-based and collaborative techniques.

3 Preservation planning and Plato

The Planets[1] project has developed a systematic approach for evaluating potential alternatives for preservation actions and building thoroughly-defined, accountable preservation plans for keeping digital content alive over time. In this

[1] `http://www.planets-project.eu`

approach, preservation planners empirically evaluate potential action components in a controlled setting and select the most suitable one with respect to the particular requirements of their institution [15]. The procedure is independent of the solutions considered; it can be applied for any class of strategy, be it migration or emulation or different approaches and follows a variation of utility analysis. The selection procedure leads to well-documented, well-argued and transparent decisions that can be reproduced and revisited at a later point in time.

The planning tool Plato [1] implements the preservation planning workflow and supports, documents, and automates the decision procedure. Following the planning process in Plato results in a well-documented preservation plan [8] one can be held accountable for. The software itself has been implemented as a JavaEE compliant web application relying on open frameworks such as Java Server Faces and AJAX for the presentation layer and Enterprise Java Beans for the backend. It is integrated in an interoperability framework that supports loose coupling of services and registries through standard interfaces and provides common services such as user management, security, and a common workspace. Based on this technical foundation, the aim is to create an interactive and highly supportive software environment that advances the insight of preservation planners and enables proactive preservation planning.

Figure 1 shows the preservation planning environment with the workflow and the relevant entities and repositories influencing the respective phases. The planning workflow that leads to the preservation plan prescribes four phases:

1. Define requirements. The first phase documents constraints and influence factors on potential preservation strategies. It then continues with a thorough description of the collection and the chosen sample objects from that collection and concludes with the definition of the complete set of requirements. At the end of this phase the planner has a detailed and exact understanding of the collection and the preservation goals. The elicitation of the institution's requirements is the core activity in the planning workflow and vital as the requirements co-determine the optimal preservation action within the institution's context. Not all university libraries or national libraries for instance share the same objectives.
2. Evaluate alternatives. The second phase starts with discovering potential preservation actions (alternatives) which are then evaluated in a quantitative way. Controlled experiments are carried out, applying the alternatives to the defined sample objects and analysing the outcomes with respect to the requirements. The result of this phase is an evidence base that underlies all decisions to be taken in the subsequent phases.
3. Analyze results. In the third phase the results of the experiments are analysed and aggregated. The result of this phase is a ranked list of alternatives whereas that alternative with the highest performance value presents the recommended preservation action.
4. Build preservation plan. In the final phase, based on the recommended preservation action a preservation plan is created which corresponds to the

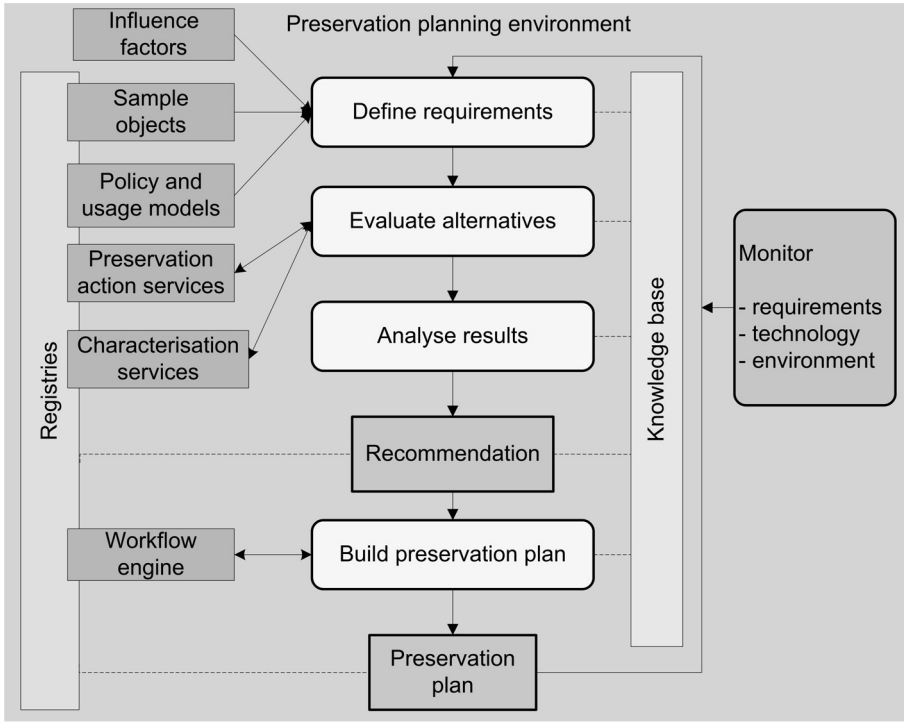


Fig. 1. Preservation planning environment

Develop Packaging Designs and Migration Plans functionality in the OAIS model [9].

Another important role in the planning process play registries holding web services for preservation action and characterisation that can be run on digital objects. While characterisation services such as DROID[2] and JHove are used to understand the digital objects on hand, action services are being evaluated and selected based on characterisation results. Furthermore the planning is supported by a knowledge base that holds reusable patterns and templates for requirements recurring in different planning situations.

4 Recommender systems for preservation planning

4.1 Introduction

The process of creating a preservation plan in Plato can benefit from recommender systems in 5 different areas:

1. Recommend particular tree template and/or fragments.

[2] `http://droid.sourceforge.net`

2. Recommend representative sample objects from a collection of objects.
3. Reduce the total amount of preservation action candidates by recommending top-N preservation action services to the user.
4. Recommend properties that can automatically be measured to be mapped to user requirements.
5. Recommend transformation templates.

The recommendations given by the system are based on different sources of information available to the recommender component. Five different sources of information have been identified so far that can be used as a basis for deriving a recommendation:

- *Usage model.* Describes in a machine interpretable way how the different users work with the collection and which priorities they have.
- *Collection profile.* Describes the characteristics of the objects in the collection and the distribution of object types within the collection.
- *Policy model.* Captures the overall organisational characteristics and requirements of the repository.
- *Knowledge base.* The knowledge database provides tree templates and fragments as well as templates for transformation table and potential preservation alternatives. The database stores the gained experience and knowledge from earlier preservation planning activities.
- *Testbed Results.* Benchmark data that allow an objective ranking of preservation action services.

Following the 5 different areas of integration are explained in more detail.

4.2 Recommend particular tree template and/or fragments

The objective tree, which is the basis of the preservation planning workflow, is usually created in a workshop setting with stakeholders from different domains. All of them contribute to the requirements gathering process. The elaborated requirements tree documents the individual preservation requirements of an institution for a given homogeneous collection of objects. [15] report on a series of case studies and describe objective trees created in these. In the stage of eliciting the requirements the preservation planner can benefit from recommendations concerning the requirements tree either in the form of templates or tree fragments. Template trees represent best practice branches and sub-trees of specific planning contexts such as for different institutions or different types of digital objects. These templates can be used as a starting point to build new objective trees as well as for refinement of existing objective trees. The templates can be adjusted for each respective preservation context. The decision which template/fragment to recommend is influenced by four entities:

- *Policy model.* Certain preservation policies can be adopted as requirement. A preservation policy which specifies the preservation strategy to follow (e.g.

migration or emulation) can be translated into a requirement and associated with a measurement scale. The same is true for policies determining an open-source strategy which can be adopted as a requirement for applied preservation actions.
- *Usage model.* The recommended requirements can be further extended including knowledge about how the various users work with the collection.
- *Chosen sample objects.* The description of the collection the preservation plan is created for and the chosen sample objects from that collection determine the content family: application, audio, video, image, or text. Each content family contains a representative pre-defined template tree that can be suggested to the planner.

4.3 Recommend Representative Sample Objects from a Collection

In the second step of the preservation planning workflow ('Define Sample Records') the planner selects sample records representing the variety of object characteristics of the considered collection. These samples are later used for evaluating the preservation alternatives. As the experiments and the evaluation of the outcome depend on the selected sample records they have to be chosen adivsedly. A comprehensive collection profile using DROID for identification and FITS[3] for metadata extraction will be created. Based on that profile the system selects a minimal set of sample objects covering a maximum number of object characteristics.

4.4 Recommend Top-N preservation action services

Discovering potential preservation actions is one of the most challinging and time-consuming tasks in the planning process. Numerous tools are available that come into question, each of them need specific input parameters and rely on a particular environment. To find relevant preservation action services the planner has to bear all these constraints in mind and rifle through existing preservation action registries. A recommender system in this stage of the workflow can reduce the amount of potential preservation action services available to the planner by recommending top-N preservation action services. The recommendation will be a ranking that is based on:

- *Migration path.* Direct migration with no intermediate conversion are preferred and thus ranked higher.
- *Collection profile.* The sample objects and the definition of the collection the preservation plan is created for determine the file format the action service must be able to handle; others can be filtered out.
- *Requirements defined by the planner and institutional policies.* For example, based on an institutional policy advising that only migration strategy shall be applied emulation services can be filtered out. This reduces the amount of

[3] `http://code.google.com/p/fits/`

candidate preservation actions and alleviates the decision which alternatives to choose for evaluation.
- *Test results.* Objective evaluation results produced by experiments carried out in the Planets Testbed[6, 2] are considered in the ranking of potential preservation actions. Actions which performed better in an experiment on comparable objects are ranked higher.

4.5 Recommend Mapping

Comparison services such as the *comparator* developed in the course of the eXtensible Characterisation Language [4, 16] specify measurable properties as well as property-specific metrics and their implementation as algorithms in order to identify degrees of equality between two objects. This is in principle independent of the applied strategy, i.e. migration or emulation. The compared objects can be both the original and a migrated object, or the original object in two different environments for emulation. To allow comparison and evaluation, a mapping is created between the requirements specified in the objective tree and the characteristics that can be measured and compared automatically by the available characterisation tools. At present this is done manually for each leaf criterion. Each leaf criterion can be mapped to a property that can be measured automatically.

The mapping done in this workflow step can be supported by a recommender system in some very basic way using pattern matching. Two possible example scenarios are given below:

- An automatically measurable property named `imageHeight` can be recommended as a candidate property to be mapped to requirements named height of image, image's height or height when it appears in a sub tree called object characteristics.
- A requirement named `Size` in a sub-tree object characteristics can be suggested to be split up into `imageHeight` and `imageWidth` because those can be measured automatically by a comparator service.

4.6 Transformation templates

Requirements are measured in different scales and are made comparable by mapping to a uniform scale using transformation tables. The resulting scale might for instance range from 0 to 5. A value of 0 denotes an unacceptable result and thus serves as a drop-out criterion for the whole preservation alternative. This transformation has to be done for each leaf criterion in the objective tree.

Figure 2 illustrates a branch, focussing on technical characteristics of the collection, which has been taken from a requirements tree of a specific institution. In a requirements tree the leaves of the tree determine the scale of the respective requirement, ranging from Y/N (either yes or no), percent of market share, to seconds per MB. The requirement 'Open specification' refers to the openness

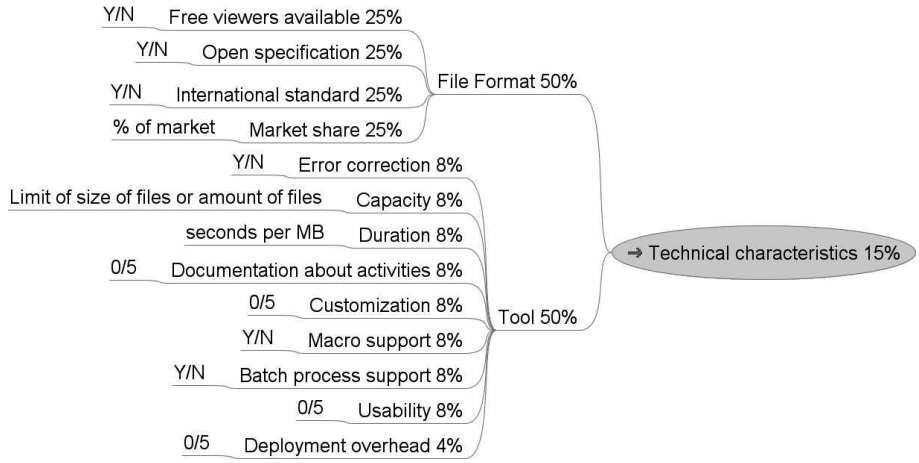


Fig. 2. Requirements referring to technical characteristics

of the file format specification which can either be 'Yes' (the specification is publicly available) or 'No' (the specification is proprietary).

Two different instituions can have two different transformation tables for the same requirements. One institution for instance can define a proprietary format specification as a drop-out criterion and another one as still acceptable. The first institution would map 'No' to zero and the second one to a value larger than zero, e.g. 1.0. The system when recommending a transformation table must consider:

- *Policy model.* The institutional policies allow the system to differentiate between acceptable and not acceptable criterions.
- *Usage model.* Similar to the policies the usage profile defines if a requirement will be defined as a drop-out criterion. The usage profile for instance declares that users perform full text search. Loosing searchability by a preservation action shall thus be not acceptable.
- *Transformation tables* established by other institutions stored in the knowledge base. The same or slightly adapted transformation tables can often be adapted for similar institutions.

5 Discussion and Outlook

This paper discusses and outlines specific integration points for recommender systems in the preservation planning tool Plato. On the one hand by this integration the effort needed to create a thorough preservation plan can be reduced. On the other hand it furthermore enables preservation planning for users with less experience in digital preservation. The next steps towards recommendation supported preservation planning are:

1. The current implementation of preservation policies in Plato is based on an extensive tree describing the policies in a structured but static way. Import

from the openly available Freemind tool is possible. In that tree a policy consists of the policy statement and a freely defined measurement scale to which degree the policy applies. However, to avoid ambiguity policies that potentially influence decisions made during the planning process have to be integrated into the system and associated with a unique identifier. This allows the recommender to use those policies for filtering results.
2. A detailed collection profile and risk analysis is vital for the planning process because the selection of potential preservation actions and the experimentation depend on it. Existing characterisation tools such as DROID and FITS will be used to extract the relevant object characteristics and create a collection profile that can then be used by the recommender.
3. Existing registries hold web service descriptions of preservation actions. Besides the technical information a WSDL contains, additional information it optionally carries, about supported data types and operations, is mainly in natural language. To enable an automated selection of preservation action services they must be enriched with QoS information and metadata such as licensing and required platform of the underlying tool. OWL-S [11], an ontology adopted for web services, may be used to create a computer-interpretable description of these web services to allow proper recommendations.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

1. *Plato: A Service Oriented Decision Support System for Preservation Planning* (2008).
2. Aitken, B., Helwig, P., Jackson, A. N., Lindley, A., Nicchiarelli, E., and Ross, S. The planets testbed: Science for digital preservation. *Code4Lib 1*, 5 (June 2008). See http://journal.code4lib.org/articles/83.
3. Balabanovic, M., and Shoham, Y. Combining content-based and collaborative recommendation. *Communications of the ACM 40* (1997), 66–72.
4. Becker, C., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. A generic XML language for characterising objects to support digital preservation. In *23rd Annual ACM Symposium on Applied Computing (SAC'08)* (Fortaleza, Brazil, March 16-20 2008).
5. Burke, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction 12*, 4 (2002), 331–370.
6. Farquhar, A., and Hockx-Yu, H. Planets: Integrated services for digital preservation. *International Journal of Digital Curation 2*, 2 (November 2007), 88–99.
7. Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry. *Commun. ACM 35*, 12 (1992), 61–70.
8. Hofman, H., Planets-PP subproject, Becker, C., Strodl, S., Kulovits, H., and Rauber, A. Preservation plan template. Tech. rep., The Planets project, 2008. `http://www.ifs.tuwien.ac.at/dp/plato/docs/plan-template.pdf`.

9. ISO. *Open archival information system – Reference model (ISO 14721:2003).* International Standards Organization, 2003.
10. Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., and Cohen, M. D. Intelligent information-sharing systems. *Commun. ACM 30*, 5 (1987), 390–402.
11. Martin, D. Owl-s: Semantic markup for web services.
12. Mirza, B. Jumping connections: A graph-theoretic model for recommender systems. Master's thesis, Computer Science Dept., Virginia Tech, 2001.
13. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. Grouplens: an open architecture for collaborative filtering of netnews. In *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work* (New York, NY, USA, 1994), ACM, pp. 175–186.
14. Resnick, P., and Varian, H. R. Recommender systems. *Commun. ACM 40*, 3 (1997), 56–58.
15. Strodl, S., Becker, C., Neumayer, R., and Rauber, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proc. of JCDL'07* (June 2007), pp. 29–38.
16. Thaller, M., Heydegger, V., Schnasse, J., Beyl, S., and Chudobkaite, E. Significant characteristics to abstract content: Long term preservation of information. In *ECDL '08: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 41–49.

Session 3

Data and text analysis

Named Entity Recognition in Biomedical Texts

Gabriel Tutoky[1], Marian Lapko[1]

[1] Faculty of Electrical Engineering and Informatics, Technical University of Košice,
Letná 9, 042 00 Košice, Slovakia
{gabriel.tutoky, marian.lapko}@gmail.com

Abstract. This paper presents usability of Named Entity Recognition (NER) in biomedical texts. After brief introductory of principles information extraction, NER and terminology issues the named NER system in practice is presented. There is an example of visual analysis of biomedical documents with visual results of analysis. In the conclusion of this paper are described results of example.

1 Introduction

Text is the predominant medium for information exchange among experts. The volume of biomedical literature is increasing at such a rate that it is difficult to efficiently locate, retrieve and manage relevant information without the use of text-mining applications.

Text-mining (TM) and also *knowledge extraction* are ways to aid researchers in coping with information overload. Originally, TM was defined as the automatic discovery of previously unknown information by extracting information from text [1]. However, in the biomedical community, the term TM is often reduced to the process of highlighting (i.e. retrieving of extracting) small nuggets of relevant information from large collection of textual data.

TM typically consists of:

- *Information Retrieval* (IR), which gathers and filters relevant documents;
- *Information Extraction* (IE), which selects specific facts about prespecified types of entities and relationships of interest [1];

2 Information Extraction

There are exist several definitions of IE [2], [1], which are more or less acceptable, one of them is by [3] following definition: Information extraction is identification and subsequent classification of specific information founded in nonstructural sources (e.g. in textual sources written in natural language) to semantic categories thus current information will become a better information for additional processing.

IE is a type of document processing which captures and outputs factual information contained within a document. Similar to IR system, an IE system responds to a user's information need. Whereas an IR system identifies a subset of documents in a large text database, an IE system identifies a subset of information within a document [2].

An IE system may perform several types of tasks; one of them is text tagging. In a text tagging task, information of particular generic type is identified, for example: all of the drug names within a collection of journal articles, or all of the gene names and symbols within a collection of MEDLINE abstracts [4]. The output of such system might be used, for example, to create indexes of documents for later information retrieval applications. Another example might be the display of the original text directly to an analyst, with relevant types of information marked or highlighted in some way [2] (see pictures below in section 4).

3 Named Entity Recognition

It is widely agreed that *Named Entity Recognition* (NER) is an important step for various applications of natural language processing such as IR, Machine Translation (MT), IE and natural language understanding [5].

The goal of NER is identification all names of entities (usually terms) of specific type of thing within a collection of textual documents; for example in biomedical domain they are names of all genes or proteins. Idea is that recognizing biological entities in text allows for further extraction of relationships and other information by identifying the key concepts to be represented in some consistent, normalized form. So NER is a crucial step for more complex IE systems.

Creation of the complex system for NER of biomedicine terms is faced to several problems, there are following types:

- *Domain and Dictionary problem*
- *Terminology of Natural Language*

3.1 Domain and Dictionary problem

Obviously, it is impossible build any system for NER which is able to cover all domains in the world. Selection of the single domain is necessary (e.g. biomedical domain) for building of trust systems. Is needed to note, that single biomedical domain is so complex and there does not exist a complete dictionary for most types of biological named entities, so simple text-matching algorithms are not suffice. Next problem is that every day hundreds of new names of proteins or genes are created.

3.2 Terminology of natural language

Main problem of terminology consist from the fact that there is often no one-to-one correspondence between concepts and terms. In practice, TM application are faced with the problems of term variation and term ambiguity

Term variation originates from the ability of a natural language to express a single concept in a number of ways. For example, in biomedicine there are many synonyms for proteins, enzymes, genes etc [1]. One of the concrete examples is following: concept name for specific antibiotic should be 'A-23187', but there are several ways to express it: 'A-23187', 'A23187', 'Antibiotic A23187', 'A23187, Antibiotic'.

Term ambiguity occurs when the same term is used to refer multiple concepts. Ambiguity is an inherent feature of natural language. Words typically have multiple dictionary entries and the meaning of a word can be altered by its context [1].

As the following problem of terminology is tokenization. It is demonstrative in gene/protein name NER because punctuation cannot be globally removed to make processing straightforward. Gene and protein names often contain hyphens, parentheses, brackets, and other types of punctuation. In [6] and [7] were presented by L. Tanabe, J. Wilbur et al. four rules applied to recognize gene and protein names in their AbGene system.

4 Named Entity Recognition in Practice

Now we will show a real example for demonstration of NER usability. Idea is displaying of the original text directly to visual analyst with highlighted relevant types of information (see section 2). We will trying to recognize and filtering all articles with relevancy to description any body part, anatomic or histological structure.

Current example can be solved in following steps:

- *Construction of relevant dictionary* – this is a crucial step; at *first*, the robust dictionary is needed; for the *second*, relevant records or categories to current task must be filtered.
- *Matching algorithm* – algorithm should have to solve more complex cases than simple matching of the words.
- *Viewing of the results* – good and user friendly output of results

4.1 Construction of relevant dictionary

At *first* - for creation of our dictionary we used two sources of biomedical terms. Both of sources are published by US National Library of Medicine:

- *Medical Subject Headings (MeSH)* – controlled dictionary for indexing MEDLINE/PubMed articles. Used version includes 25 186 key words classified to 133 semantic categories. The vocabulary contains concept names and its possible variation e.g. concept name – Floor of Mouth;

possible terms – Floor, Mouth; Floors Mouth; Mouth Floor; Mouths Floor; etc. [9]

- *Unified Medical Languag System (UMLS)* – syntactic Lexicon of biomedicine. Used version includes 386 746 records. The Lexicon contains set of acceptable expressions for various terms e.g. term – mouth; acceptable expressions – mouths, mouthed, mouthing [10]

For the *second* – we investigated both sources of biomedical sources. First source – MeSH dictionary was used as a main source of connections between terms and its meanings. In MeSH all of the terms are assigned to one or more semantic categories. We extracted all terms for each category from it and we studied these categories with biomedicine expert. He jointed similar semantic categories into one dictionary e.g. we created one dictionary with relevant terms to any body part, anatomic or histological structure from ten particular categories of MeSH like *Cell*, *Body System*, *Anatomical Structure*, etc. Second source – UMLS Lexicon was used for extension of dictionaries obtained from MeSH with new variations of existing terms.

4.2 Matching algorithm

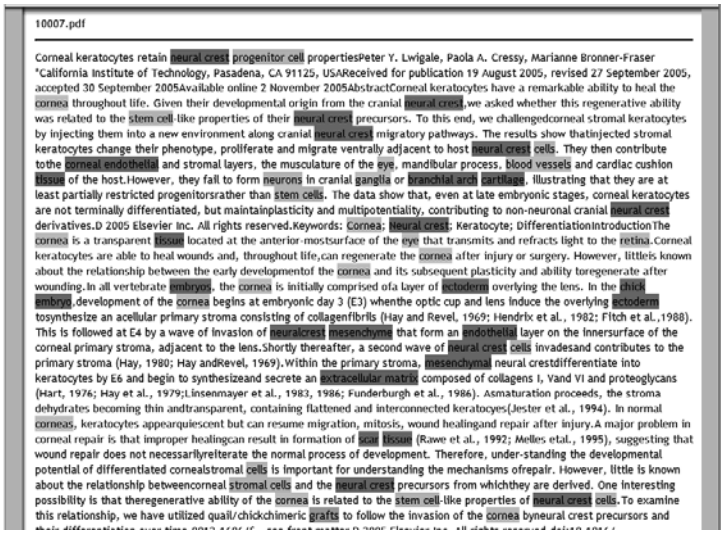
Matching algorithm is very important component of NER systems. By its quality are dependent results of the system, especially measures like precision and recall. Also effectiveness and computational complexity of algorithm is important, but with more robust algorithm its complexity is increasing.

The following cases should be solved by matching algorithm:

- *Simple matching of single terms*
- *Simple matching of multiple terms* – some possible options:
  - Number of white characters could be various
  - No strict punctuation characters between terms, e.g. dot "." vs. comma "," (in Slovak, the real numbers are written with coma, but in English with dot)
- *Case sensitive/insensitive*
- *Identifying similar terms* by computing an distance between them, e.g. Levenshtein edit distance with restriction to term length
- *Hyphenated terms* – usually text extractors extract hyphenated words from various formats with dash

4.3 Viewing of the results

For viewing of results we used simple web page, where all relevant (founded) terms are highlighted with some color. The same color means same semantic category of terms. Few types of results are depicted in pictures below.



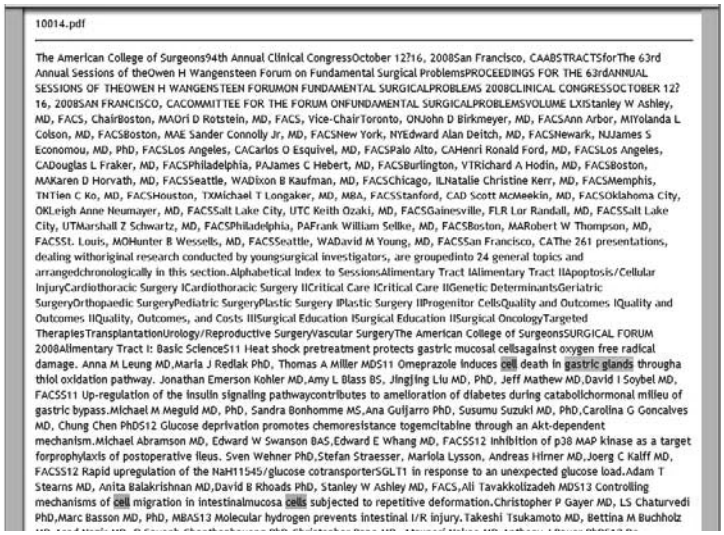
10007.pdf

Corneal keratocytes retain neural crest progenitor cell propertiesPeter Y. Lwigale, Paola A. Cressy, Marianne Bronner-Fraser
*California Institute of Technology, Pasadena, CA 91125, USAReceived for publication 19 August 2005, revised 27 September 2005, accepted 30 September 2005Available online 2 November 2005AbstractCorneal keratocytes have a remarkable ability to heal the cornea throughout life. Given their developmental origin from the cranial neural crest,we asked whether this regenerative ability was related to the stem cell-like properties of their neural crest precursors. To this end, we challengedcorneal stromal keratocytes by injecting them into a new environment along cranial neural crest migratory pathways. The results show thatinjected stromal keratocytes change their phenotype, proliferate and migrate ventrally adjacent to host neural crest cells. They then contribute tothe corneal endothelial and stromal layers, the musculature of the eye, mandibular process, blood vessels and cardiac cushion tissue of the host.However, they fail to form neurons in cranial ganglia or branchial arch cartilage, illustrating that they are at least partially restricted progenitorsrather than stem cells. The data show that, even at late embryonic stages, corneal keratocytes are not terminally differentiated, but maintainplasticity and multipotentiality, contributing to non-neuronal cranial neural crest derivatives.D 2005 Elsevier Inc. All rights reserved.Keywords: Cornea; Neural crest; Keratocyte; DifferentiationIntroductionThe cornea is a transparent tissue located at the anterior-mostsurface of the eye that transmits and refracts light to the retina.Corneal keratocytes are able to heal wounds and, throughout life,can regenerate the cornea after injury or surgery. However, littleis known about the relationship between the early developmentof the cornea and its subsequent plasticity and ability toregenerate after wounding.In all vertebrate embryos, the cornea is initially comprised ofa layer of ectoderm overlying the lens. In the chick embryo,development of the cornea begins at embryonic day 3 (E3) whenthe optic cup and lens induce the overlying ectoderm tosynthesize an acellular primary stroma consisting of collagenfibrils (Hay and Revel, 1969; Hendrix et al., 1982; Fitch et al.,1988). This is followed at E4 by a wave of invasion of neuralcrest mesenchyme that form an endothelial layer on the innersurface of the corneal primary stroma, adjacent to the lens.Shortly thereafter, a second wave of neural crest cells invadesand contributes to the primary stroma (Hay, 1980; Hay andRevel, 1969).Within the primary stroma, mesenchymal neural crestdifferentiate into keratocytes by E6 and begin to synthesizeand secrete an extracellular matrix composed of collagens I, Vand VI and proteoglycans (Hart, 1976; Hay et al., 1979;Linsenmayer et al., 1983, 1986; Funderburgh et al., 1986). Asmaturation proceeds, the stroma dehydrates becoming thin andtransparent, containing flattened and interconnected keratocytes(Jester et al., 1994). In normal corneas, keratocytes appearquiescent but can resume migration, mitosis, wound healingand repair after injury.A major problem in corneal repair is that improper healingcan result in formation of scar tissue (Rawe et al., 1992; Melles etal., 1995), suggesting that wound repair does not necessarilyreiterate the normal process of development. Therefore, under-standing the developmental potential of differentiated cornealstromal cells is important for understanding the mechanisms ofrepair. However, little is known about the relationship betweencorneal stromal cells and the neural crest precursors from whichthey are derived. One interesting possibility is that theregenerative ability of the cornea is related to the stem cell-like properties of neural crest cells.To examine this relationship, we have utilized quail/chickchimeric grafts to follow the invasion of the cornea byneural crest precursors and

Figure 1. Article with high frequency of terms from analyzed domain

reduced in Bmpr2 H11546 / H11546 embryos. To evaluate how the lack of BMPR2 affects BMP signaling in the embryo, we examined the distribution of phosphorylated Smad1/5 (p-Smad1) in Bmpr2 H11002 / H11002 embryos by immunohisto-f_luorescence staining. In wild-type embryos, p-Smad1 was found in the nuclei of epiblast ,and primitive endoderm cells at E4.5 and of epiblast and VE cells at E5.2 (Fig. 1 N and Fig. 2, A and B). [ID] The distribution of p-Smad1 changed quickly between E5.2 and E5.5 and had shifted to the proximal epiblast and VE, ex-cluding DVE, at E5.5 (Fig. 2 C and Fig. S3, A ? J, available at http://www.jcb.org/cgi/content/full/jcb.200808044/DC1). In Bmpr2 H11002 / H11002 embryos, the distribution of p-Smad1 was similar to that in wild-type embryos at E4.5 but showed two distinct pat-terns at later stages (Fig. 1 N and Fig. 2 , A ? ? C ?). In severely af-fected mutant embryos (8/14 embryos at E5.2 and 8/15 embryos at E5.5), p-Smad1 was apparent only in the proximal VE at E5.2 (8/8 embryos; Fig. 2 B ?) and was barely detected at E5.5 (8/8 embryos; Fig. 2 C ?). In mildly affected embryos (6/14 embryos at E5.5 and 7/15 embryos at E5.5), p-Smad1 was found in the same regions as in wild-type embryos at E5.2, but its abundance was lower than that in the wild type (6/6 embryos; Fig. 1 N and Fig. S2). It was not detected in the epiblast and there were fewer positive cells in the VE of the mildly affected embryos at E5.5 Results DVE formation is impaired in Bmpr2 H11546 / H11546 embryos Formation of the primitive streak is impaired in Bmpr2 H11002 / H11002 em-bryos (Beppu et al., 2000). To determine whether formation of the A-P axis occurs normally in these mutant embryos, we f_lexamined the expression of AVE or DVE marker genes at E6.5 and E5.5, respectively. In wild-type embryos at E5.5, f_i DVE marker genes, Lefty1 , Cerl , Dkk1 , Lim1 , and Hex , are expressed in VE at the distal tip (Fig. 1, A ? E). [I Expression of Hex , Hesx1 , and Cerl is absent at E5.2 but is apparent at E5.5 (Fig. S1, A ? C and E ? G, available at http://www.jcb.org/cgi/content/full/jcb.200808044/DC1), whereas Lefty1 expression is maintained be-tween E4.0 and E5.5 (Takaoka et al., 2006 ; Fig. S1, D and H), indicating that cells positive for a full range of DVE markers are formed between E5.2 and E5.5. In Bmpr2 H11002 / H11002 embryos, how-ever, expression of AVE marker genes at E6.5 was absent or re-duced compared with that in wild-type embryos (Fig. S2, A ? D, A ? ? D ? , and A ? ? ? D ? ?). Dkk1 expression was lost (Fig. S2 C ?) or remained relatively normal (Fig. S2 C ? ?). At E5.5, expression of Lefty1 , Cerl , Dkk1 , and Lim1 was absent (4/7, 3/7, 3/7, and 3/6 embryos, respectively) or markedly reduced (3/7, 4/7, 4/7, and 3/6 embryos, respectively), and that of Hex was also lost (3/3 embryos; Fig. 1 , A ? ? E ? and N; and Fig. S2, I and I ?). To determine the region of the embryo in which BMP signaling exerts the observed effects, we examined expression of DVE marker genes in green embryonic stem (ES) FM260 cell ? ? Bmpr2 H11002 / H11002 tetraploid chimeric embryos, which were generated by aggregation of ES cells expressing EGFP with Bmpr2 H11002 / H11002 tetraploid embryos. In such chimeras, expression of Hex (n = 3) and Lefty1 (n = 3) was absent at E6.5 (Fig. 1 , F ? H and F ? ? H ?). This phenotype was indistinguishable from that of Bmpr2 H11002 / H11002 embryos, suggesting that BMPR2 in the extra-embryonic region is required for DVE formation. We next examined whether VE is formed normally in Bmpr2 H11002 / H11002 embryos. VE, which is composed of embryonic VE and extraembryonic VE at E5.5, is derived from the primitive endoderm of the E4.0 ? 4.5 embryo. Expression of Hex , which is a marker of the primitive endoderm, was maintained in Bmpr2 H11002 / H11002 Figure 2. p-Smad1 in wild-type and Bmpr2 H11546 / H11546 embryos. Wild-type (A ? C) or Bmpr2 H11002 / H11002 (A ? ? C ?) embryos at the indicated stages of development were subjected to immunohistofluorescence staining with antibodies to p-Smad1 (pS1; green); merged images with staining of nuclei by YOYO-1 (Nuc; red) are also shown. Staining for p-Smad1 was decreased in Bmpr2 H11002 / H11002 embryos. Bars, 50 μ m. on January 27, 2009 jcb.rupress.orgDownloaded from JCBp-Smad1 staining was observed in green ES FM260 cell ? ? Bmpr2 H11002 / H11002 ,Actr2b+/ H11002 tetraploid chimeric embryos (Fig. S3, K and L). These results suggested that both BMPR2 and ActR2b act as receptors for BMP in VE. We next examined DVE markers in Bmpr2 H11002 / H11002 ,Actr2b+/ H11002 embryos at E5.5 to determine whether BMP signaling is required for DVE formation. Expression of DVE markers was detected in some of the Bmpr2 H11002 / H11002

Figure 2. Article with high frequency of one or few terms from analyzed domain, in figure these terms are "*embryo*" and "*endoderm*".


10014.pdf

The American College of Surgeons94th Annual Clinical CongressOctober 12?16, 2008San Francisco, CAABSTRACTSforThe 63rd Annual Sessions of theOwen H Wangensteen Forum on Fundamental Surgical ProblemsPROCEEDINGS FOR THE 63rdANNUAL SESSIONS OF THEOWEN H WANGENSTEEN FORUMON FUNDAMENTAL SURGICALPROBLEMS 2008CLINICAL CONGRESSOCTOBER 12? 16, 2008SAN FRANCISCO, CACOMMITTEE FOR THE FORUM ONFUNDAMENTAL SURGICALPROBLEMSVOLUME LXIStanley W Ashley, MD, FACS, ChairBoston, MAOri D Rotstein, MD, FACS, Vice-ChairToronto, ONJohn D Birkmeyer, MD, FACSAnn Arbor, MIYolanda L Colson, MD, FACSBoston, MAE Sander Connolly Jr, MD, FACSNew York, NYEdward Alan Deitch, MD, FACSNewark, NJJames S Economou, MD, PhD, FACSLos Angeles, CACarlos O Esquivel, MD, FACSPalo Alto, CAHenri Ronald Ford, MD, FACSLos Angeles, CADouglas L Fraker, MD, FACSPhiladelphia, PAJames C Hebert, MD, FACSBurlington, VTRichard A Hodin, MD, FACSBoston, MAKaren D Horvath, MD, FACSSeattle, WADixon B Kaufman, MD, FACSChicago, ILNatalie Christine Kerr, MD, FACSMemphis, TNTien C Ko, MD, FACSHouston, TXMichael T Longaker, MD, MBA, FACSStanford, CAD Scott McMeekin, MD, FACSOklahoma City, OKLeigh Anne Neumayer, MD, FACSSalt Lake City, UTC Keith Ozaki, MD, FACSGainesville, FLR Lor Randall, MD, FACSSalt Lake City, UTMarshall Z Schwartz, MD, FACSPhiladelphia, PAFrank William Sellke, MD, FACSBoston, MARobert W Thompson, MD, FACSSt. Louis, MOHunter B Wessells, MD, FACSSeattle, WADavid M Young, MD, FACSSan Francisco, CAThe 261 presentations, dealing withoriginal research conducted by youngsurgical investigators, are groupedinto 24 general topics and arrangedchronologically in this section.Alphabetical Index to SessionsAlimentary Tract IAlimentary Tract IIApoptosis/Cellular injuryCardiothoracic Surgery ICardiothoracic Surgery IICritical Care ICritical Care IIGenetic DeterminantsGeriatric SurgeryOrthopaedic SurgeryPediatric SurgeryPlastic Surgery IPlastic Surgery IIProgenitor CellsQuality and Outcomes IQuality and Outcomes IIQuality, Outcomes, and Costs IIISurgical Education ISurgical Education IISurgical OncologyTargeted TherapiesTransplantationUrology/Reproductive SurgeryVascular SurgeryThe American College of SurgeonsSURGICAL FORUM 2008Alimentary Tract I: Basic Science511 Heat shock pretreatment protects gastric mucosal cellsagainst oxygen free radical damage. Anna M Leung MD,Maria J Redlak PhD, Thomas A Miller MDS11 Omeprazole induces cell death in gastric glands througha thiol oxidation pathway. Jonathan Emerson Kohler MD,Amy L Blass BS, Jingjing Liu MD, PhD, Jeff Mathew MD,David I Soybel MD, FACS511 Up-regulation of the insulin signaling pathwaycontributes to amelioration of diabetes during catabolichormonal milieu of gastric bypass.Michael M Meguid MD, PhD, Sandra Bonhomme MS,Ana Gutjarro PhD, Susumu Suzuki MD, PhD,Carolina G Goncalves MD, Chung Chen PhDS12 Glucose deprivation promotes chemoresistance togemcitabine through an akt-dependent mechanism.Michael Abramson MD, Edward W Swanson BAS,Edward E Whang MD, FACS512 Inhibition of p38 MAP kinase as a target forprophylaxis of postoperative ileus. Sven Wehner PhD,Stefan Straesser, Mariola Lysson, Andreas Hirner MD,Joerg C Kalff MD, FACS512 Rapid upregulation of the NaH11545/glucose cotransporterSGLT1 in response to an unexpected glucose load.Adam T Stearns MD, Anita Balakrishnan MD,David B Rhoads PhD, Stanley W Ashley MD, FACS,Ali Tavakkolizadeh MDS13 Controlling mechanisms of cell migration in intestinalmucosa cells subjected to repetitive deformation.Christopher P Gayer MD, LS Chaturvedi PhD,Marc Basson MD, PhD, MBAS13 Molecular hydrogen prevents intestinal I/R injury. Takeshi Tsukamoto MD, Bettina M Buchholz

Figure 3. Article with low frequency of terms from analyzed domain.

5 Conclusions

Presented results - visual analysis of articles by NER system (pictures above) should be classified into few categories:

- *Relevant articles* – the frequency of relevant terms is high
- *Narrow specialized articles* – the frequency of relevant terms is high, but highlighted terms are of the same semantic category (usually one or few different terms are marked)
- *Irrelevant articles* - the frequency of relevant terms is low or none
- *Articles written in another language* – no relevant terms

In nowadays TM applications become excellent assistant for biomedical research to obtain useful information in still shorter time. TM supplications so have high potential in biomedical domain, which is not realized yet. It is necessary to put pressure to cooperation with TM and biomedical researchers during application development.

Acknowledgement

The work presented in this paper was supported by the following projects: the Slovak Research and Development Agency under the contract Nr. RPEU-0011-06; the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant Nr. 1/4074/07; Slovak Ministry of Health project GEMIN under Nr. 2007/65/UPJŠ-02.

References

1. Spasic I., Ananiadou S., McNaught J., Kumar A.: Text mining and ontologies in biomedicine: Making sense of raw text. In Proceedings of the Briefings in Bioinformatics, vol. 6 no. 3, Pages 239-251, Henry Stewart Publications 1467-5463, September 2005.
2. Okurowski, M. E.: Information Extraction overview. In Proceedings of the Workshop held at Fredricksburg, Pages 117-121, Virginia, September 19-23, 1993.
3. Moens, M.-F. Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series); Springer: 2006.
4. Cohen A. M., Hers W. R.: A survey of current work in biomedical text mining. In Proceedings of the Briefings in Bioinformatics, vol. 6 no. 1, Pages 57-71, Henry Stewart Publications 1467-5463, March 2005.
5. Sassano M., Utsuro T.: Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In Proceedings of the 18th conference on Computational linguistics - Volume 2, Pages 705-711, Saarbrücken, Germany, 2000.
6. Tanabe L., Wilbur J.: Tagging gene and protein names in biomedical text. National Center for Biotechnology Information, NLM, NIH, Pages 1124-1132, Bethesda, Maryland 20894, USA, February 14, 2002.
7. Tanabe L., Xie N., Thom L. H., Matten W., Wilbur J.: GENETAG: a tagged corpus for gene/protein named entity recognition. In Proceedings of the A critical assessment of text mining methods in molecular biology, Granada, Spain, March 28-31, 2004.
8. Feldman, Ronen and Sanger, James: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007
9. US national Library of medicine: *Medical Subject Headings (MeSH)*. On-line at: http://www.nlm.nih.gov/mesh/. 2009
10. US national Library of medicine: *Unified Medical Language System (UMLS)*. On-line at: http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. 2009

Information Extraction about Citation Networks

Martin Repka[1]

[1] Faculty of Electrical Engineering and Informatics, Technical University of Košice,
Letná 9, 042 00 Košice, Slovakia
repkam@post.sk

Abstract. This paper is dealing with introduction to Citation Networks and their analysis. It contains definitions of basic types of Citation Networks and divisions. It provides an overview about current and available methods and ways in Citation Network Analysis. It introduces problems of graph theory and solving graph-based algorithms. It practically uses methods of common network analysis to demonstrate a Citation Network Analysis. It shows some types of workflows and approaches to solve a large networks analysis.

Introduction

Citation networks is subclass of more general class Bibliographic networks, they can be sometimes referred in bibliographic sources as Publication Citation Networks or Paper Citation Networks.

Information extraction about citation networks can be divided into some ways. The simplest way to provide information extraction is to consider bibliographic network as a directed graph and provide modified graph analysis. Briefly, it will consist from three approaches in network analysis: vertex weights based on centralities, main path analysis based on arc weights as proposed in work of and ranking algorithms based on PageRank.

Generally, a publication cites many other publications, and those cited publications are listed as the bibliography (or references) at the end of the paper. In a citation network, a publication corresponds to a vertex (node) and citation relation between two papers corresponds to an edge (link). So in citation network the actors (set of units) are publications and the relation of citing is directed and means inverse relation to relation “cited by”.

Bibliographic Networks

Bibliographic Networks denominates general network containing bibliographic data such as publications citation, collaborations between authors or authorship itself. Due the kind of relations and items present in the network we can define this simple division:

1. Citation Networks (Publication Citation Networks or Paper Citation Networks)

2. Co-authorship Networks.

Relations in Bibliographic Networks

1. *Citing relationship*
2. *Co-authorship*
3. *Authorship*

Citing Relationship is relation present in Citation Networks. It determines connection between two publications *p,q* from set of publications *P* ($p,q \in P$). Relation $p\Re q$, $\Re \subset P \times P$ means that publication *p* cites publication *q*. It is obvious that inverse relation exists and it means *q* is cited by *p*.

In Collaboration Networks the specific relation is **co-authorship.** Formally $\Re \subset A \times A$ is relation on set of authors *A* and defines relationship between authors $a,b \in A$, which is a fact that author *a* has at least one common work with author *b*. The fact that common work can be measured (number of common papers, amount of common research etc.) this relationship can be valued.

Third kind of relation **authorship** in the base stone to mentioned relations. It defines connection between two sets of units (set of publication *P* and set of authors *A*) by expression $\Re \subset A \times P$, and $a\Re p$ means that author *a* is an author of the publication *p*.

Standard form of Citation Network

In some approaches of citation network analysis such as main path analysis uses extended form on citation network called citation network in standard form. To achieve this extended networks the transformation is needed. The set of units *U* is extended by adding special vertices *s,t* (common source, common terminal).Then corresponding edge from *t* to *s* (also called "arc") is added.

This extension to standard form eliminates problems with networks with several components or several initial or terminal units. It connects all possible initial units to one unit (source *s*) and all possible terminal units to one unit (terminal *t*). This transformation destroys network acyclicity.

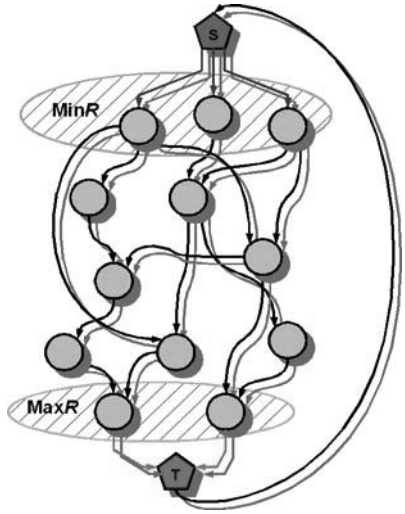


Figure 1 Citation network in Standard form

Ways of Citation Network Analysis

Citation network analysis can be made using two main ways. First way is to evaluate vertices (nodes) in networks. For this analysis can be used topological order of the vertices in near neighborhood or from global view. The most common analysis using vertices is investigating measures of centralities. In undirected cases of networks (as could be collaboration networks) we speak about measures of centralities and in directed case (as the citation networks are) about measures of importance (influence and support).

Another way is to investigate edges (connections or links) in citation network using geodesic evaluation. In most cases citation weighs analysis in standard form of Citation network is used, introducing indexes such as node pair projection count and others.

Vertex weights and Edge weights

The simple network analysis for vertices from mathematical graph theory is provided by centrality measures. This measures uses topological information about position of vertex to determine relative importance of a vertices. Centrality is structural attribute of vertices in network. More “made-to-measure” way is to use an special method for analysis as methods Hubs and Authorities and modified ranking algorithms (CiteRank etc.) The inspiration for ranking algorithms is definitely PageRank. In the time of his appearance many similar or inspirited algorithm were presented such as similar algorithm for determining important web pages called HITS

or also called Hubs and Authorities. Other ranking methods and modifications were quickly developed SALSA, SCEASRank, ObjectRank, BackRank, AuthorRank, etc. They were modified to success as methods for rating special individuals in networks like citation networks are.

For citation weights we mentioned methods as main path analysis. Main path Analysis is used for investigating connectivity in acyclic networks. This way of analysis is interesting when vertices are item dependent. In a citation network can be direction of edge given by time and each vertex can be a distinct event in time. Main-path analysis is suggested not to be applied to cyclic networks were vertices can belong to paths that lead back to themselves. A vertex with high indegree and high outdegree will probably be part of the main path. Purpose of revealing of main path is to find those publications that build on prior publications and are referenced to later publications.

In citation network a degree measure considers the number of incoming citations for a document (indegree) and number of cited references in the documents (outdegree). Then the main path is constructed by selecting those connected publications with the highest scores until an end document is reached.

By calculating this scores, main-path algorithms enable us to make the structural backbone of a literature visible. There are three models to identify the most important part of a citation network: the Node Pair Projection Count, the Search Path Link Count and the Search Path Node Pair.

Node Pair Projection Count weights are one of the indexes of edge connectivity based on the measure of traversal counts in search paths through the network. It uses the subgraph of the network, that represents all possible paths for vertex pair. If there are n vertices in the subgraph, there exist $n*(n - 1)$ possible subgraphs connecting all directed vertex pairs in the network. Then the traversal counts for each edge is computed using the adjacency matrices for all the subgraphs connecting these vertex pairs. The traversal counts of interest are the projected counts of all edges connecting vertex pairs projected onto a base matrix. The resulting projection matrix contains counts of the number of times each edge was involved in connecting all vertex pairs using all subgraphs derived from the network. We call this the node pair projection count (NPPC) method of generating traversal counts.

CNG Manager as Tool for Analysis

CNG Manager is my own implemented application to support Citation Network Analysis. It is implemented in programming language JAVA and runs platform independent. It uses package Layout Pro developed by JGraph Ltd for visualization and JGraphT package for graph data structure. Program itself has to provide an first overview on unknown Citation Network. User has his research field and he needs a point to start and a leash for moving further in publication. This is especially important, if the user is orientating in publications.

As important publications could be suggested publications with high rank scored by ranking algorithms or a centrality. On the visualized output it could be distinguished by color or scale. As the leash, where to continue in exploring CN,

could be a main path based on the weighted edges (citations). After processing and analysis, the graph output could be reduced to only important publications and citations. This step improve in big importance readability of output meaning.

Program takes xml file containing Citation Network as input it is parsed to JAVA object DefaultCNGraph using XmlOutPutAdapter. CnaCore is main part of system, doing main job (Citation network analysis), where all methods for Citation Networks Analysis are implemented such as Network flow analysis, Graph Centralities, Ranking methods and Main Path Analysis. It handles all change-invoking operations on DefaultCNGraphT Object. Graph Layout maintains network visualizing and interpreting results of analysis. In fact it is a layout manager derived from layout manager JGraph Layout Pro designed by JGraph Ltd. for providing automatic positioning of the graph, network or diagram accordingly to general layout rules.

Program Workflow

Program can be used in basic workflows. User decides which workflow will be used to proceed analysis with CNG Manager.

Simple workflow is straight workflow, in first step citation network is loaded from source (xml file) then suitable analysis is proceeded and results are optionally used for output reduction. As "suitable analysis" in vertices investigation you can choose a variety of centrality methods or ranking method PubRank For edges' analysis you can choose NPPC weights or main path analysis.

To analyze a big network geodesic traversal methods (eg. based on the shortest paths) are very expensive, because mainly it is needed to investigate all the shortest path from any to any vertex in network. So we can use **complex workflow** in CNG Manager. Inexpensive method to investigate all vertices is taken to determine, which of them are important and which are not. The methods such as local centralities or eigenvector centrality based ranking algorithms are usually inexpensive and quick. The best candidates for quick analysis of vertices are: Degree Centralities, Hubs and Authorities and PubRank. As we have analyzed vertices, we can decide with vertices are not useful anymore. Then we make reduction in graph and we create graph consisting of important vertices. On reduced network we can run more expensive methods as geodesic based centralities for vertices and main path for edges are.

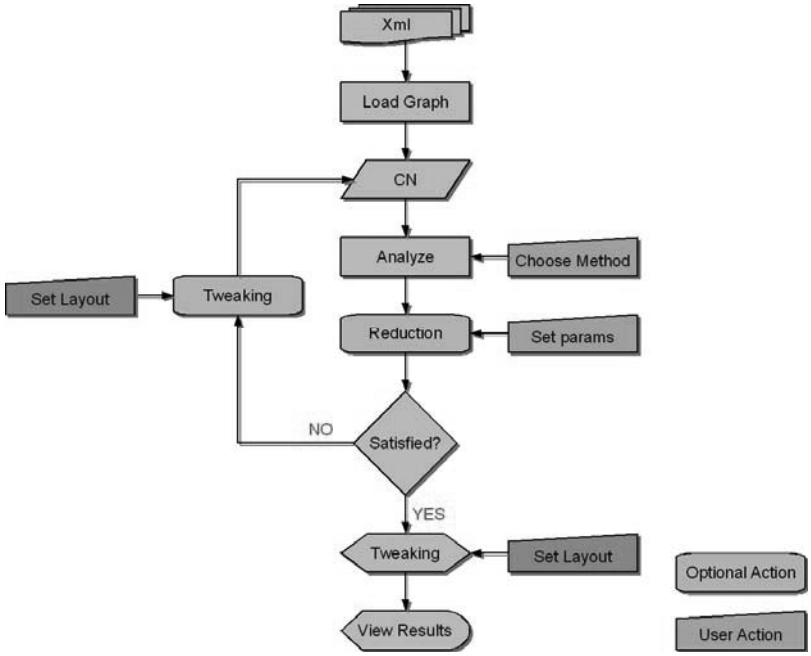


Figure 2 Complex workflow in CNG Manager

Analysis Interpretation in Bibliographic content

Indegree Centrality reflects indegree of all vertices in graph, in Citation Network using citing relation it means, that the most cited publications are highly scored. These articles are usually very important in the research field and are possibly older than most of other publications. If we use inverse relation of citing relation in Citation Network , this feature “most cited” will reflect the outdegree centrality. **Outdegree centrality** measures outdegree of vertices in graph. As we use standard citing relation the centrality measures the publications, which cites a large number of other publications. Publications with that feature are usually some kind of resume, reviews or overviews in the certain field. If we use inverse relation this feature “cites a large number of publications" will reflect the indegree centrality.

Next methiod for analysis that detects hubs and authorities called as **Hubs and Authorities**. The publications with high hub score serves as large directories. They are not very authoritative, they usually represent compilations of broad catalog of information. These publications connect directly to other authoritative publications and otherwise to the good authoritative publications many different hubs are linked.

PubRank uses eigenvector based PageRank on Citation Networks. It rates publications like PageRank ranks web pages. It is useful analysis, because the rank is determined after several iteration and it is not based only topological position of publication in network, but it consider also ranks from surrounding publications.

Node pair projector count is method based on traversing geodesics in the network. In Networks weights the edges by counting lengths of incoming paths to source vertex and lengths of outgoing paths from target vertex. It ranks edges along the path, which leads from authority to fresh articles with minimal score, it is strongly dependent in structure of Citation Network, which should be time-depend and hierarchic. In Citation Networks is this method used for revealing the main path or backbone of research. It also reveals the narrow parts of the Citation networks

5 Experiment

This experiment demonstrates practical use of program by user. They all are provided on the example of Citation Network consisting of 2692 publication from medical research field and 5178 citiations. It consist from one method of analysis.

This example shows using of PubRank scores to determine 100 most important vertices and building an reduced network with information about geodesics remained.

After loading citation network it appeared as unreadable amount of vertices and edges.Then method PubRank was chosen and graph was reduced to 100 most important vertices (due to PubRank). Then NPPC analysis of edges was run and the edges were weighted. The vertex cells are scaled and colored due to PubRank, blue and the biggest ones are with the biggest PubRank score.

Resulting graph has 100 vertices and 504 edges created in process \ref{reduction}. PubRank is from range 0.188 to 1.0 and edges are 1.0 to 6.0 long. On this graph we made reductions to 75 edges .

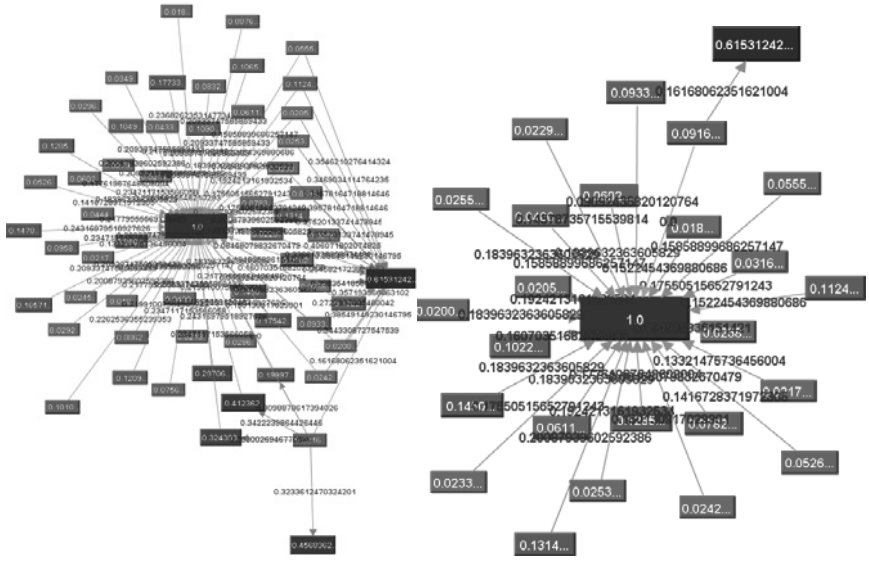


Figure 3 Reduced output to 75 edges and final output (25 edges)

On the figure we can see, that remaining vertices are 4 highly ranked by PubRank (probably most important articles in network), which are blue and violet. And the less important publications, which cite it. A lot of them cite the most ranked publications. As we proceed to next reduction (75 edges) we can see publications connected between two most ranked publications (1.0 and 0.61).

As we proceed the final reduction to 25 edges we obtain a graph with 26 vertices. The final output is an most rated vertex by PubRank and its surroundings. The most ranked vertices could be considered as suggestion to start exploring in this citations.

6 Conclusion

This paper briefly presents approaches in citation network analysis and described developed experimental tool. This method of analysis seem to be suitable for large citation network analysis. All algorithms have been inspired or are derived from existing and efficient systems.

Acknowledgement

The work presented in this paper was supported by the following projects: the Slovak Research and Development Agency under the contract Nr. RPEU-0011-06; the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant Nr. 1/4074/07; Slovak Ministry of Health project GEMIN under Nr. 2007/65/UPJŠ-02.

References

1. Sugiyama, K., Ohsaki, H., Imase M.: Structural Analysis of Paper Citation and Co-Authorship Networks using Network Analysis Techniques, Graduate School of Information Science and Technology, Osaka University, Japan, 2006.
2. Bagatejl, V.: Course on Social Network Analysis Weights. Padova, April 2003, University of Ljubljana
3. Bagatejl, V.: Efficient Algorithms for Citation Network Analysis. Ljubljana, September 2003, University of Ljubljana
4. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry, 1977
5. Borgatti, S.: Centrality and network flow. New Orleans: Sunbelt International Social Networks Conference, 2002
6. Jezek, K., Fiala, D., Steinberger, J.: Exploration and Evaluation of Citation Networks. Toronto: June 2008, ELPUB2008 Conference on Electronic
7. Maslov, S., Redner, S.: Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks. Journal of Neuroscience 28,
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project, 1998

Log-based Analysis of Knowledge Processes

Jozef Wagner[1], Ján Paralič[1]

[1]Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics Technical University of Košice, 042 01 Košice, Slovakia
jozef.wagner@tuke.sk, jan.paralic@tuke.sk

Abstract. This paper describes proposed technical framework for log-based analysis of knowledge processes that are performed within a virtual user environment. Each user action or activity is monitored and stored in the activity log in the central repository. Before storing, activities are translated according to the reference model ontology, suitable for subsequent analysis. Query interface is provided on the top of the log, and is used by various analytical tools which can be divided into two types: analytical tools which present summarized information about performed actions and tools which visualize processes as sequences of actions on a timeline.

1 Introduction

Virtual collaborative systems typically provide functionalities and possibilities to realize user actions and activities in a collaborative manner in order to create new knowledge based on transformation of existing different sources. The second important goal of these systems is to teach users how to work in a collaborative environment with utilization of created relations between them or between them and functional environment. It is necessary to monitor and further investigate performed activities in order to identify positive or negative aspects of applied solution, to identify strong or critical points in whole process, etc. This type of analyses is covered by proposed technical framework and theoretical background behind it.

In most real world scenarios, where the information system supports collaborative activities, the processes that happen between users are hard to describe. The traditional process theories or even theories around business processes hardly cover what is going on in the system. And if we take into the account the activities which where performed outside the boundaries of the information system, we find that the processes are becoming very complex. The so called knowledge processes involve humans as well as information systems. The inputs and outputs of these processes are not clearly defined.

Teacher-student scenario, a course, can serve as a real world example of such processes. Within this course, a collaborative work on complex issues is taking place. Students participate on a collaborative development and enhancement of knowledge objects. The assignments are open-ended, and based on this the whole working structure is ill defined. Being the real world course, it is only partly supported by an

Information system, as many activities can take place in the field, not in the information system, thus the system is unable to capture it automatically.

Such activities can often be of much importance to the knowledge process, can serve as an interesting practice and can be subject to further analysis. Examples of such field activities include:

- Interactions: Transcribed recordings of meetings
- Reflections: Teachers' reflections in diaries, semi-structured interviews with students and teachers;
- Knowledge objects: Reports, meeting notes, concept maps, and handwritten comments

Analytical tools developed within our information system are used by the researchers as well as by regular users – students and teachers. The main goals behind the functionalities of the analytical tools are to:

- examine team and individual activities as well as development of knowledge objects during the course
- increase teachers understanding of knowledge creation process
- review history of knowledge creation
- support for reflection on the knowledge objects
- provide basis for teachers intervention

1.1 Related work

A framework for analyses and visualizations of collaborative processes was designed within the Kaleidoscope project[1]. This framework, called CAViCoLA (Computer-based Analysis and Visualization of Collaborative Learning Activities) provides functionalities to identify existing complex interactions within examined processes [3]. The analysis results are visualized in an appropriate graphical format that enables users to make their own interpretations and allows them to reflect on their previous activities.

Several different analytical approaches can be found in domain of business processes that can be found as quite similar to mentioned type of collaborative processes. Interesting initiative in this are is called process mining with possibility to extract of potentially useful information from event logs. Event logs are results of monitored activities that are implemented within execution environment for business processes. ProM [1] represents a generic open-source framework for implementing process mining tools in a standard environment. The ProM framework receives as input logs in the Mining XML format (MXML[2]) and currently, this framework has plug-ins for process mining, analysis, monitoring and conversion.

[1] http://www.noe-kaleidoscope.org/pub/

[2] http://en.wikipedia.org/wiki/MXML

3 Architecture

Proposed technical framework for log-based analysis of knowledge processes, called History and Participation Awareness service (HPA) is designed and developed within IST European project called Knowledge Practices Laboratory (KP-Lab). HPA is a middleware service which receives from KP-Lab tools a set of events to be stored in the log repository (internally, the data is stored in the MySQL database). As each tool has its own application domain model according to which it works with data, the event received from the tool is then translated according to the reference model ontology. Lastly, the HPA provides a set of query interfaces for the analytical tools, in order to present historical data back to the user. The translation to the reference model enables to present this data in a unified and comprehensive way. Fig. 1 provides a schematic look on the part of the system architecture involving HPA.

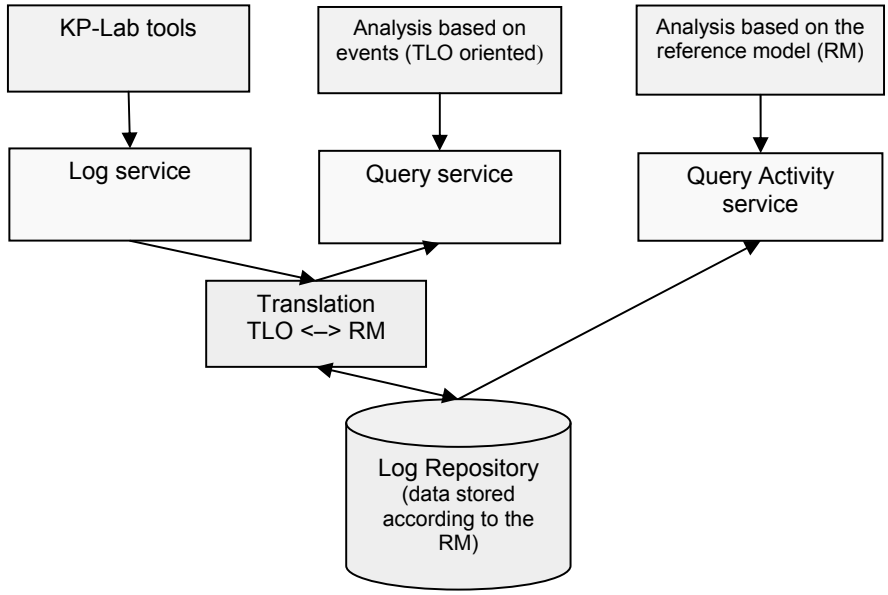


Fig. 1. Architecture of the awareness service

The Reference Model ontology is an ontology developed within KP-Lab project in order to bridge a gap between a technical and pedagogical understanding of the knowledge processes running in the information system. The Reference Model reflects expressions of common abstractions and conceptualizations drawn from the empirical data, reflected as descriptive, high-level, comprehensive and extensible open sets of specializations. Unlike tools ontologies, the Reference Model correctly preserves the time aspects of knowledge processes, by focusing on the activities performing across the information system. The Fig. 2 visualizes the core part of the Reference Model ontology.

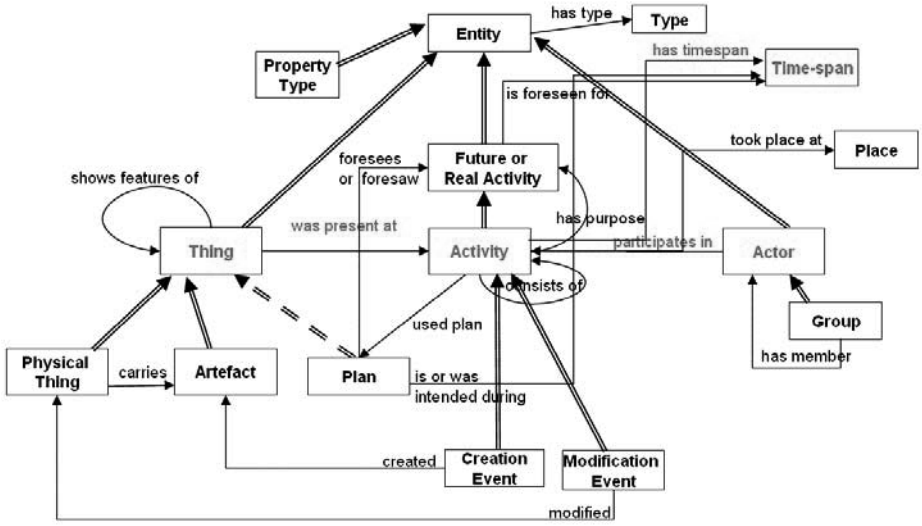


Fig. 2. Core part of the Reference Model ontology

3.1 Query interface

By having all tool events stored in a central repository and in a common format according to the Reference Model, HPA is able to provide a unified interface on top of the logs. It is mainly aimed for analytical tool asking for specialized aggregated information that will further be processed (visualized, used for decision support etc.). The following web service provides a simple aggregation view:

```
String activityAggregation (Query query,
    List<AggregationFunction> aggregationFunctions,
    Set<GroupBy> groupBy)
```

Query – this parameter describes constraints which will be used for filtering of the activities included in the aggregated view. Query object encapsulates the following constraints already specified for HPA:

- activityType - type of performed activity,
- entityID - URI of the Object of activity,
- actorID - URI of the Actor,
- timeRange - time interval,
- objectBelongsTo – ID of the object to which entity
- filter - set of key value pairs which will be compared with events custom properties,
- excludeFilter - true or false, whether include or not events which does not have properties from the filter present in them.

aggregationFunctions: specify the list of aggregation functions included in the view computed from the set of selected events.

- NumOfEvents - the number of events,
- NumOfActors - the number of unique subjects/actors included in the result,
- NumOfEntities - the number of unique objects/entities included in the result,
- TimeSpan - the date of the first event and date of the last event (starting and ending date).

groupBy: specify clause for the grouping of the result. It is possible to specify the following values:

- actor - group results by the subject,
- entity - group results by the object,
- activityType - group result by the type of the activity.

return value: XML of the result

Example. This example will present aggregated view, which will select all users working on the Task1 object and for each user it will contain the number of updates and time span when the user updated this object:

```
group by actor, Query(entityID = Task1, activityType =
update), aggregationFunctions = NumOfEvents, TimeSpan
```

Result:

```
<result>
 <row>
  <actor>User1</actor>
  <numOfEvents>10</numOfEvents>
  <startingDate>10-11-2008</startingDate>
  <endingDate>20-11-2008</endingDate>
 </row>
<row>
  <actor>User2</actor>
  <numOfEvents>2</numOfEvents>
  <startingDate>10-11-2008</startingDate>
  <endingDate>12-11-2008</endingDate>
 </row>
</result>
```

For simpler queries, following service is provided:

```
public Activity[] getRecentActivities(TimeRange timeRange,
    Property[] filter, String excludeFilter, int from,
    int max)
```

filter - set of key value pairs which will be compared with events custom properties. Moreover, following keys can be used:

- actorId, actorType, actorName
- entityId, entityType, entityTitle
- activityType
- time, objectBelongsTo

excludeFilter - true or false, whether include or not events which does not have properties from the filter present in them

from, max – if user does not want all results, (s)he can limit query results to those that fall within a specified range. *From* parameter specifies starting point, and *max* parameter specifies maximum number of events/activities returned. If *from* parameter is negative, the results are returned in reverse order.

Results returned by this service are Activities. These classes are simple Java Beans and are further described in HPA documentation[3].

3.2 Analytical tools

Analytical tools are integrated part of user virtual environment within KP-Lab project. They provide functionalities to visualize results of user queries, to visualize sequences of actions based on user expectations, to import external events, to create user patterns and customization of visualization based on user level of experiences or access role.

First analytical tool outlined in this paper is the Timeline-based analytical tool, which provides the users/researchers with a graphical interactive user interface presenting the selected group of events from the awareness repository on a timeline (Fig. 3). Besides simple timeline, the following functionalities are provided:

- The users are able not only to preview the selected sequence of events representing various activities by their types, but also to inspect particular events in detail as well as to particular associated documents.
- Moreover, the user is able to insert information about external events that were relevant to analyzed process, also e.g. referring to physical artefacts or actors outside the information system etc.
- Important functionality of this tool is the annotation support according different schemes, and lastly, the ability to define and store selected patterns (pattern is a set of selected - usually annotated – events) from the timeline.

[3] http://kplab.tuke.sk/hpa

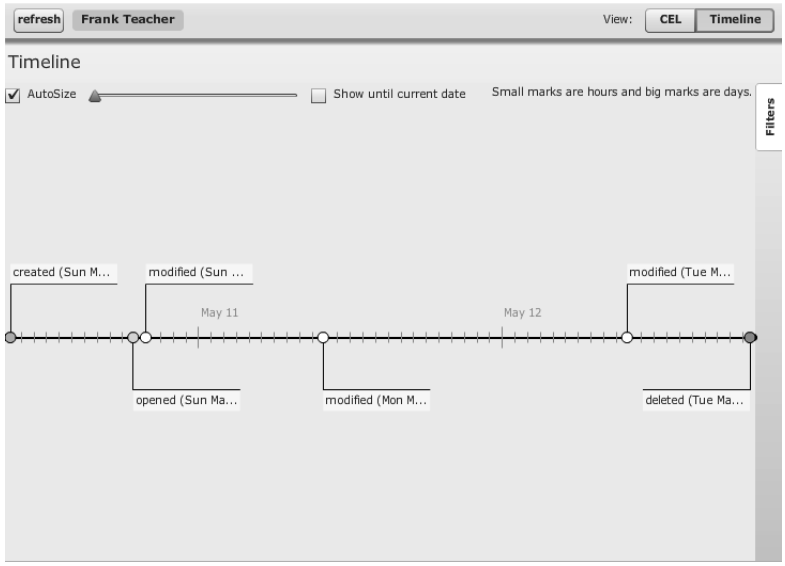


Fig. 3. Actual screenshot of time-line GUI

Besides timeline-based analytical tool, KP-Lab information system also has a Visual analyzer, which provides means for selecting and visualizing a subset of activities logged in awareness repository. User are assisted by this tool in definition of various statistical queries that will be performed by Analytic and Knowledge Mining Services and the results in form of summarized data are exposed in suitable visual form of presentation (Fig. 4)

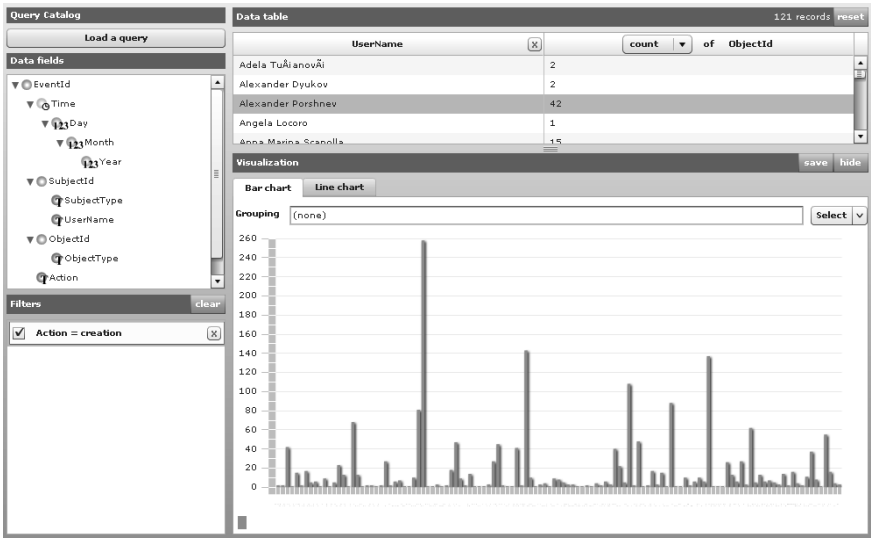


Fig. 4. A visual analyzer

4 Conclusion

Proposed technical framework for monitoring and analyzing of collaborative processes within virtual user environment is a work within IST European project called KP-Lab. Awareness service presented in this paper unified the management of events performed within different tools in the information system. Common query interface on top of the stored logs allowed analytical tools to work consistently with all knowledge processes that took place. Actual version of designed solution, mainly the analytical tools, is still in development phase, but several partial prototypes are already available for testing purposes within field trials and pilot courses. Results of these tests will be used for improvement and development of stable release that will be available in January 2010 as an integrated part of KP-Lab System[4].

Acknowledgment

The work presented in this paper was supported by: European Commission DG INFSO under the IST program, contract No. 27490; the Slovak Research and Development Agency under the contract No. APVV-0391-06 and RPEU-0011-06; the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/4074/07.

The KP-Lab Integrated Project is sponsored under the 6th EU Framework Programme for Research and Development. The authors are solely responsible for the content of this article. It does not represent the opinion of the KP-Lab consortium or the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

1. Dimitriadis, Y.: Computer-base Analysis and Visualization of Collaborative Learning Activities (CAViCoLA). Kaleidoscope Symposium 2007: Defining the Scientific Evolution of Technology Enhanced Learning", Berlin, December 2007.
2. Alves de Medeiros, A.K., et al. Semantic Process Mining Tools: Core Building Blocks. In W. Golden, T. Acton, and K. Conboy, editors, 16th European Conference in Information Systems (ECIS) 2008, CD-ROM, ISBN 13:978-0-9553159-2-3.
3. Babič, F., Wagner, J.: Awareness service based on events logs. 2nd Workshop on Intelligent and Knowledge oriented Technologies 2007, Košice, Slovakia. pp. 106-109, ISBN 978-80-89284-10-8.

[4] http://www.kp-lab.org/tools

Session 4

Service-oriented architecture

Automated web service composition

Zoltán ĎURČÍK

Department of Cybernetics and Artificial Intelligence, FEI TU Košice, Slovak Republic
zoltan.durcik@tuke.sk

Abstract. Web service composition is in present very discussed. Web services are programs located in networks, which may be used by users or by software agents via standard protocols. Web services communicate by XML messages, most often by SOAP messages. Each of web services provides some functionality (e.g. translation words from one language to other language). In the case if isn't possible fulfill request by one web service (e.g. we need translate word from English language into Spanish, but we haven't available service, which this directly translate), there is a possibility try to use web services composition (e.g. we assume, that we have a web service, which translates English word to German word, and next we have second web service, which translates German word into Spanish). Planning techniques, which may be used to web services composition, are e.g. graph-oriented planning, heuristic-planning, planning by logical programming etc.

Key words: web services, WSDL, OWL-S, composition, AI planning, planner

Introduction

With an increase number of web pages it devotes increasingly attention to relevant information searching problem on internet. Main problem was, that majority web pages and data on internet were oriented to user. Therefore the computer resources (e.g. software agents) had information understanding problem on internet. Hence start the discussion how assure, that software resources understand the content of internet too. The result of this discussion has been introduction a metadata into web pages hidden content. In this manner began start semantic web, which besides user is oriented to software resources (as agents), which are able to understand its content. These resources are then able effective work with the content of web, and e.g. offer relevant answers to people requests. In relation to semantic web were introduced, or has been obtain more attention, several standards and technologies. Among them belong e.g. RDF, RDF schema, OWL, OWL-S and likewise. Together with progressive evolution of internet was increasingly attention devoted to web services.

Web services are distributed programs, which are located on networks (most frequently on internet) and are used by standard protocols (most frequently by HTTP) [1]. This concept was introduced by main IT Corporation as Microsoft, IBM and Sun. Web services communicate with their user and with another web services by XML messages through Internet. Operations description, web service properties and messages format are available through web service interface. To web service

description is used a specific description language, most frequently WSDL. Main key to web services understanding is understand their standards and protocols. Between most fundamental belong WSDL - Web Service Description Language, SOAP - Simple Object Access Protocol and UDDI - Universal Description, Discovery and Integration.

WSDL is a descriptive language based on XML technology. It serves to web services description and has two main goals. First goal is describe web service and the second goal is localize web service. SOAP is XML based protocol, which serves to information exchange via network for web services, most frequently through HTTP. SOAP is a communication protocol, which serve to interaction via internet. UDDI is a standard to registration, categorization and searching web services. Method of working with UDDI reminds a catalogue, in which are saved information about web services providers and about their web services. OWL (Web Ontology Language) serves to publication and sharing ontologies. It comes from RDF (Resources Description Framework). RDF is a system to resources description on internet. Given description consists of trinity subject - predicate - object. OWL-S comes from OWL and it is ontology to web services description. OWL-S web service description consists from service profile, which describes what web services makes and what functionality provides, next from process model, which describes how web services communicates with clients, and from grounding, which specifies the ground properties of service as communication protocols, messages format, port type and likewise.

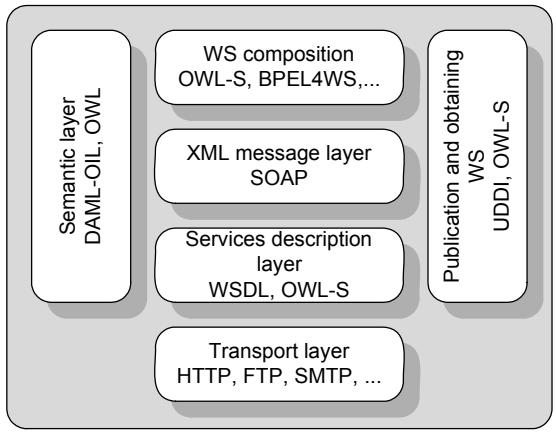


Fig. 1. A survey of presented web services technologies

One of the major problems in relation to web service and semantic web is their composition. Each of web services provides certain functionality for us (e.g. words translation from English to Slovak language). Necessity of web services composition occur then, when isn't possible fulfill request by one web service, but it is possible by several services (e.g. we need translate words from Chinese to Slovak language, but we haven't to this one service. Here are but two web services, from which first makes translation from Chinese to English, and second makes translation from English to Slovak).

As example of possible web services composition using maybe introduce the traveling domain. Traveling problem has several parts, e.g. necessity working trip to fixed date, to certain city, choose the mode of transport and likewise.

- **History** - complicated access to the information. Here was a necessity to detect e.g. telephone number to hotels, forward go to train station and buy train ticket and so.
- **Present** - on-line hotels reservations, traveling tickets, but also other events.
- **Future?** - enter query (e.g. hotel; Prague; 3.10.2009 to 8.10.2009) and the system should be able automatic find most suitable answer for user (requester), at which after result confirmed by user system could automatic executing all bank operations too (e.g. hotel payment, ticket payment and likewise).

Among other potentially domains for utilization web services belong for example medical domain, automatic responding to e-mail, text processing and text mining domain, document processing.

1. Automatic web service composition

Nowadays increasingly organization, corporation but also individuals implement their applications, or offer their services, just by web services technology. Web services have in general specific inputs and outputs. It means that after our query they provide some information following input date for us. In generality is valid that if for our query don't exist one web service, which fulfill our request, is maybe possible fulfill query by combination (composition) particular web services. A draft of web service composition system is displayed on Fig. 2.

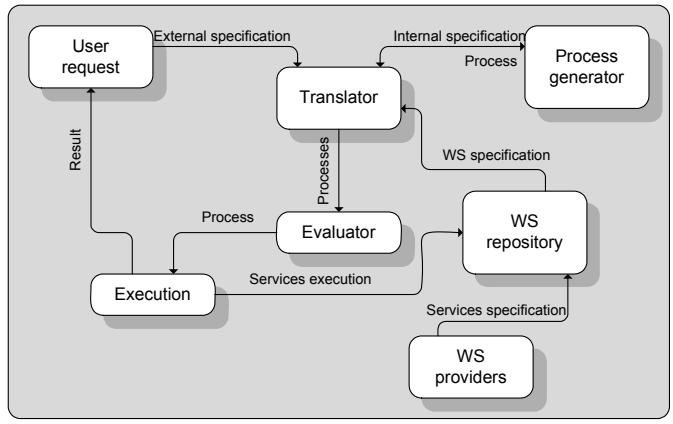


Fig. 2. Framework for web services composition

- **Service repository** - serves to storing web services obtained from service providers. It is possible e.g. through UDDI standard.

- **Translator** - is a component, which serves to information processing obtaining from users and from web services repository. In many web services composition system is language by which user enter request and a language specification, which is used by algorithm to composition, different. For example users most frequently use standard as OWL, OWL-S etc., whereas program may works e.g. on logical programming principles and uses language as Prolog.
- **Process generator** - is a main component of system. It is a set of algorithms, which choose atomic web services following the user request (query), which fulfill this request. At the end of this process is for us provided a set of atomic web services together with data and work flow among these web services.
- **Evaluation** - occur in the case when are generated several different plans. It's evident, that by current number of web services here may exists a lot of web services, which have similar functionality. Therefore if is proposed several plans, evaluator determine, which of these plans is chosen following independent values (e.g. user constraints).
- **Execution engine** - is an execution web services in order chosen by planner. Result is provided to user.

2. Artificial intelligence planning methods

As one of most suitable methods for web service composition were shown to be artificial intelligence planning methods. The planning is a solution of planning problem, when is available information about actions and about their combinations and the task is find solution over the set of all available plans. Planning problem may be represent as a world model and it is possible write its as pentad: $<S,S_0,G,A,\Gamma>$.

S represent a set of all possible states in given model, S_0 is a subset of S and marks initial state, G marks goal state, A is a set of available actions, where each of these actions change the world state by passing its from one state to another state, and a relation Γ is subset of $S \times A \times S$ and define precondition and effect for each action.

The relation between planning and automated web services composition is following: sets S_0 and G represent initial and goal state, which may be represented by ontologies, e.g. by OWL. A is a set of actions and is represented by set of available atomic web services. Γ represent a state change function for each service. As most suitable description for web services in relation to artificial intelligence planning methods is used *OWL-S* description. By OWL-S description is possible besides input and output describe precondition and effect too.

In generality planning problem consists from following part:

- description **available actions**, which may be executed,
- description of **initial state**,
- and description of **goal state**.

Among most frequently artificial intelligence planning techniques used belong state-space planning, graph-oriented planning, planning by using hierarchical task networks and planning by using logical programming.

2.1. State-space planning

State space consists from following parts [3]:

- **S** - enclosed set of states,
- **A** - enclosed set of actions,
- **f** - function, which describes transitions among individual states,
- **c(a,s)>0** - determine the "price" of application action a in state s.

State space, which includes initial state S_0 and goal state S_G description, is marked as state model.

A goal of state space model is to find a actions sequence $\{a_0, a_1, ..., a_n\}$, which generate a sequence of state $s_0, s_1=f(s_0, a_0), ..., s_{n+1}=f(s_n, a_n)$, where is valid, that is possible use action a_i in state s_i, $a_i \in A(s_i)$, where $A(s_i)$ is an enclosed set of available actions in state s_i and state $s_{n+1} \in S_G$ is a goal state.

In fact to state space oriented searching may be used random searching algorithm. State space interprets the situations from real world but may be too much extensive. Therefore we must be by the choice of searching algorithm very careful, and here it is necessary care at its performance and how it is able to manage with searching space extensity.

By the first attempts to reduce state space belonged **STRIPS** algorithm (STandford Research Institute Problem Solver) [3]. STRIPS uses backward chaining. It works with following elements:

- **initial state**
- **goal state**
- set of **actions**, where each action includes:
  - **preconditions** - they inform us about it, what must be fulfilled for executing given action,
  - **effects** (or postconditions) - they inform about it, what will be changed in the model after executing given action.

Mathematically we maybe define STRIPS as set $<A, O, I, G>$, where:

- A is a set of conditions, which are often represented as atoms
- O is a set of operators (actions), which is given as tetrad $<\alpha,\beta,\gamma,\delta>$:
  - α inform us about it, which conditions must be true for executing given action
  - β inform us about it, which conditions must be false for executing given action,
  - γ are conditions, which will be true after executing action
  - δ are conditions, which will be false after executing given action.
- I is initial state of model, which is given as set of conditions, which are in this state true and at the same time is valid, that all other conditions are considered as false. Is valid that I is subset of A.
- G is goal state of model, which is given as couple $<N, M>$. This couple specifies, which conditions are true and which are false in given state. Is valid that G is subset of A.

The relation between STRIPS and Web services is evident. Web services maybe represent as actions, which have preconditions and effects. This representation of web services simplest we get by using OWL-S for web services description. From given description is then possible obtain actions, for which belong specific precondition before executing and specific effects after executing action.

2.2. Graph oriented planning

Graph oriented planning for its activity uses graph structures. There graph structures are often marked as planning graphs [5][9]. Planning graph is different from state space graph, which represent states as graph nodes and edges as particular transitions between individual states. Planning graph consists of two types of nodes, concretely actions node and conditional nodes. These nodes are localized in alternated layers. Condition layer is following with action layer etc. This is showed on fig. 3. Here we have tree operations (Oper.1 until 3) and an initial state represented by two conditions (Con. 1 and Con. 2). The creation of graph continue until, while aren't two consecutive layers identical. From fig. 3 is evident, that by final conditional layer could be possible use operation 3. But by its using we would again obtain the same conditional layer.

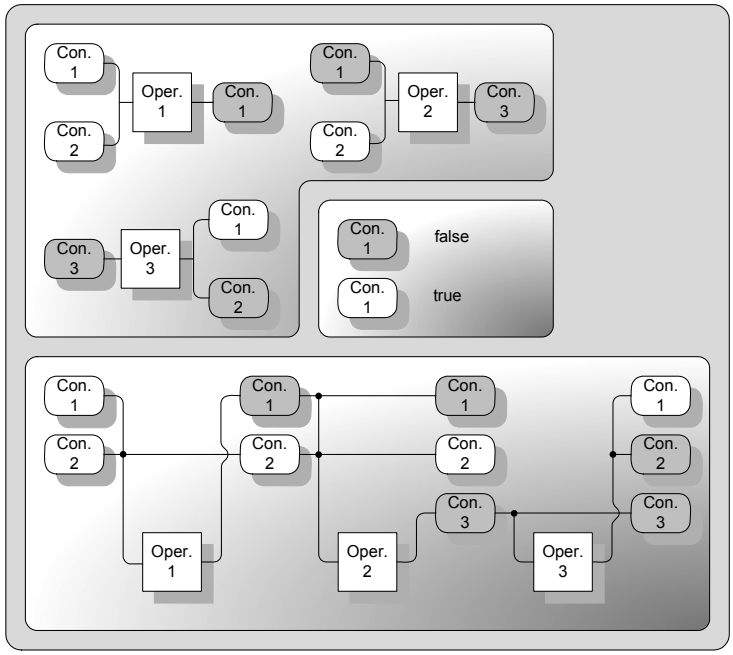


Fig. 3. Graph expansion in GraphPlan

2.3. Planning by using hierarchical task networks

Relations among particular actions of planning problem are expression by networks. Planning problem is specified in hierarchical task network by following set of possible task:

- **primitive tasks** - they respond STRIPS actions,
- **composite task** - they may be composite from several simpler tasks (it means that from primitive or from another composite tasks),
- **goal tasks**, which respond to STRIPS goal.

Primitive tasks represent actions, which may be directly executing. Composite task represent a sequence of actions and goal tasks represent conditions. For composite tasks execution must be known their sequence of primitive tasks, from which they are composite. Goal tasks represent conditions, which must be true in goal state.

A variant of hierarchical task network, which obtains the most attention, is planning by decomposition ordered tasks. Planners founded on this approach, as e.g. SHOP (Simple Hierarchical Ordered Planner) [6], accept goals as task list where composite tasks may consist of another tasks or of primitive tasks. The system of ordered task decomposition doesn't plan directly the achievement of defined goal, but it plan executing (composite or primitive) action.

The using hierarchical task network planning in web services area was showed in system SHOP2 [6], which belong into planners family based on ordered task decomposition. It present a transformation method from OWL-S processes into hierarchical task networks. Likewise as hierarchical task network OWL-S processes have predefined actions description to task executing. This description we make usually manual.

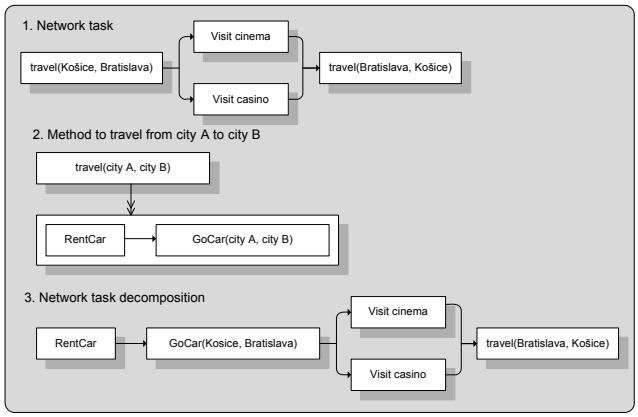


Fig. 4. Network task decomposition

On Fig. 4 we have showed decomposition of network task travel(A, B) by method RentCar and GoCar.

2.4. Planning by using logical programming

Logical programming is next approach for planning problem representation and from this founded possibility for web services composition solving. By logical programming we consider a program, which may be represented as set of Horn clauses in implication form $A \leftarrow (B_1 \text{ a } B_2 \text{ a } ... \text{ a } B_n)$. Each such Horn clause may be represented as literal disjunction (in mathematic literal marks atomic formula, thus atom, whereby positive literal presents atom and negative literal presents atom negation), where may be positive only one literal, i.e. $A \text{ v } \neg B_1 \text{ v } \neg B_2 \text{ v } ... \neg B_n$.

In a case of using logical programming for planning were showed as the most suitable methods for this the methods based on deductive reasoning, e.g. in the case of Prolog [10]. Among another logical programming application belongs e.g. Reiner implementation of Golog and situation calculus [11]. Author is oriented to knowledge oriented Golog programs, which may contain sensing actions. These programs refer to agent knowledge and they are designed to online execution under enclosed world assumption.

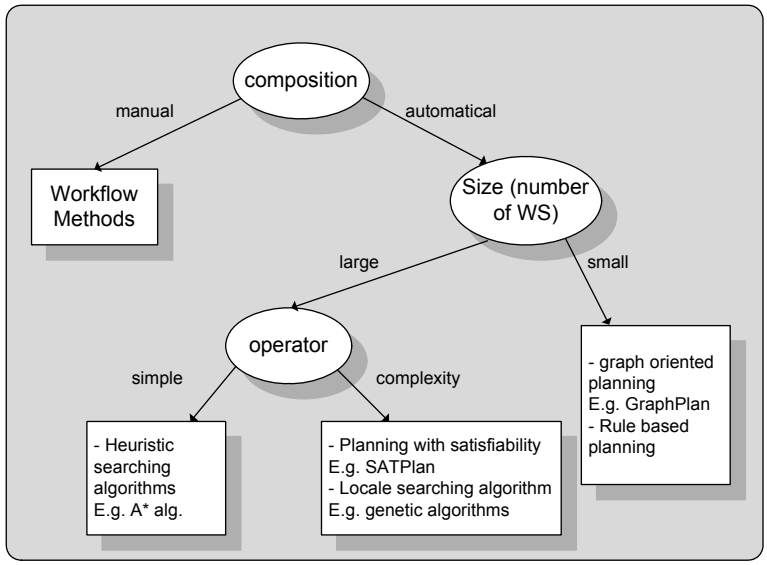


Fig. 5. Decision tree of methods for web services composition

2.5. Comparison of particular methods

On Fig. 5 we have displayed decision tree, which may be helpful by choice correct approach to web services composition following tree decisions:

- **type of composition**
  - In a case of **manual composition** we usually choose workflow methods, where is composition made manual, at which is defined the order of

services a data flow among services. These techniques it possible apply only with small number of services,
    - In a case of work with larger number of services and the necessity **automatic** web services **composition** we choose artificial intelligence planning.
- **size** of planning **domain**
    - **large** size domain
        - simple operators - use heuristic searching algorithms
        - large complexity of operators - use planning with satisfiability,
    - **small** size domain - use graph oriented planning methods, rule based planning or planning based on partial ordering.

3. Actual system for web services composition

3.1. (Semi)automatic web services composition using PROLOG

There is used an elements of logical programming, concretely Prolog language [10]. A prototype of this system has two main components (Fig. 6):

- **composition tool** (CT) and
- **inference tool** (IT).

Inference tool stores information about well known services in a knowledge base and in a case of necessity it is able to find suitable service. Composition tool makes besides composition also interaction between human operator and inference and composition tool.

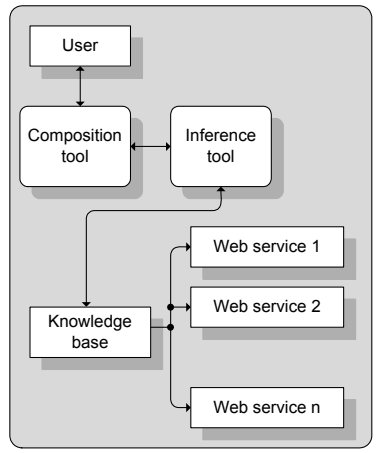


Fig. 6. System architecture for web services composition based on logical programming

Inference tool serves to OWL reasoning and is built on Prolog. Information described in OWL are converted into RDF triplet and loaded into knowledge base. Inference tool has inbuilded inference rules for OWL. These rules are applied on facts in knowledge base in order to find all dependencies. Like this is e.g. discovery heredity between two classes, which isn't directly encoded into class relations. Composition tool supports user by workflows creation by offering available services in every step. User starts composition process by choosing one registered service. Then is request sending into knowledge base in order to obtain information about service input, and next is for each input generated new request for purpose of obtain new services list, which are able to ensure given inputs. Composition tool also show different service classes available at the system and filter out result following restrictions, which were specified by user following service attributes.

3.2. Web services composition using SHOP2 planner

SHOP2 [6] is domain independent hierarchical task network (HTN) planning system, which won one from four main prices among 14 planners on international planning competition in the year 2002 (IPC-2002). HTN is artificial intelligence planning method, which is focused on plan creation by tasks decomposition. Task decomposition is performanced until they aren't all tasks decomposed onto primitive tasks.

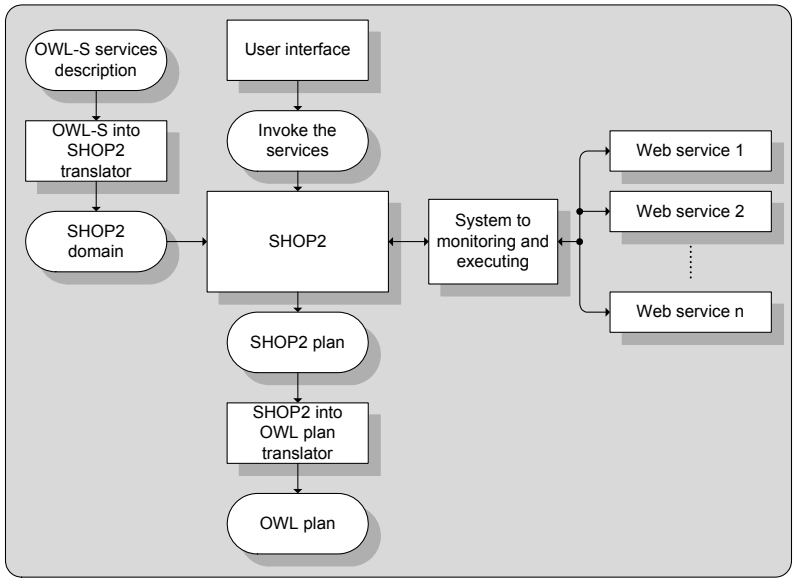


Fig. 7. SHOP2 system architecture [6]

On fig. 7 is illustrated system architecture for web service composition using SHOP2 planner. This system is composite from following main parts:

- **SHOP2 planner** - following initial state, goal state and available action creates plan, which is described in SHOP2 domain.
- Tool for **web services management** - has on task storing information about availability web services and plan execution.
- Tool for **resultant plan representation** in OWL ontology - provides plan representation in OWL standard.
- Tool for **translation OWL-S** service description **into SHOP2** domain - translates OWL-S services descriptions into SHOP2 domain (e.g. atomic processes from OWL-S into HTN planning procedures)

3.3. OWL-SXPlan

OWLS-XPlan [7] is a tool developed on web service composition realization by artificial intelligence planning method. It realizes converting web services described by OWL-S 1.1 standard into equivalent problem realized in PDDL 2.1 language. PDDL (Planning Domain Definition Language) is attempt for descriptive language for planning domain and problem standardization. It was developed for ICP (International Planning Competition 1998/2000). After this problem transcribed into planning domain is realized planning by AI planner XPlan (Fig. 8). Output from given planner is services sequence (actions) marked also as plan. XPlan is based on extension fast forward planner (FF planner) by HTN planning.

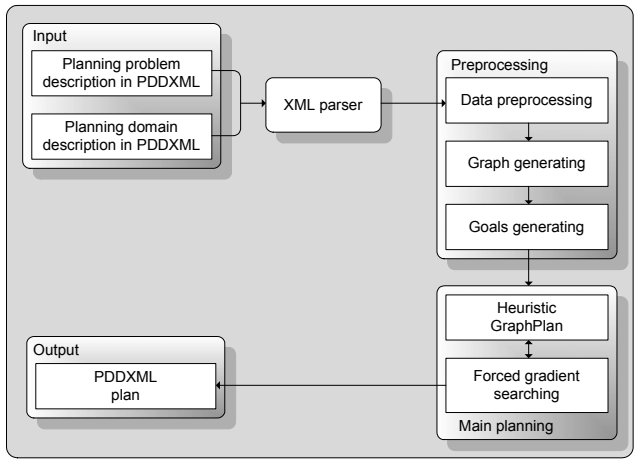


Fig. 8. XPlan planner

4. Summary AI planning methods and systems

System	Planning methods	WS standard support	Internal problem representation	System type	Dynamically interference during planning
WS composition Prolog	Logical programming	OWL-S WSDL	Logical programming program	Semiautomatic composition with user collaboration	No
Shop2 system	HTN planning	OWL-S WSDL	PDDL 2.1 version 2	Automatic composition	No
OWL-SXplan	Extended Fast forward planning FF and HTN planning	OWL-S WSDL OWL	PDDL 2.1 version 2	Automatic composition	Yes, OWL-Sxplan version 2

Table 1. Comparison of chosen systems for WS composition

Majority of presented AI planning methods have roots in STRIPS. STRIPS may be defined as set $<A, O, I, G>$, where A is a set of conditions, O is a set of operations (actions), I is initial state and G is a goal state in planning model. A goal of planning is transpose system (model) from initial state I into goal state G by operators from O and by compliance conditions from A. Here is clearly visible analogy with web service composition. In a case of composition goes likewise about planning. Initial state is a state, in which we are now. Goal state is a state, in which we would like to be. Conditions result from domain, in which we plan and from conditions, which are given for particular web services. Web services represent actions. Therefore is possible after suitable modifications and transformation use for web service composition just AI planning methods. Given transformation inhere e.g. in transformation initial and goal state, which may be described in OWL ontology, and web services described by OWL-S and WSDL into PDDL language representation. In this manner it was solved e.g. in OWLSXplan system.

The problem of composition maybe divides minimum into four parts:

- system interaction with user
- planner for web services composition
- web services obtaining
- execution obtained plan from web service composition

5. System proposal for web services composition

On fig. 9 we have presented architecture, which should have our planning system. Current is it only architecture draft and particular blocks from this draft may vary over time into final version, which will to be thereafter implemented.

Proposed system consists of following four main parts:

- **User interface** - has on care interaction between user and system. It is divided into two parts:
  - o *user part* - serves to prevalent users, which use system,
  - o *expert part* - serves to experts, which may modify system functionality, e.g. domain knowledge modification.
- **Preprocessing** - serves to creation *initial* and *goal state*, which are obtained from users. We assume that these two states will be represented by ontologies. It cooperates with domain knowledge.
- **Planner** - following initial state, goal state, available web services and information about antecedent *composition* attempts is created planning problem a planning model. This model is subsequently solved.
- **Localization and execution web services** - has on care *interaction with web services*, web services indexation and their execution.

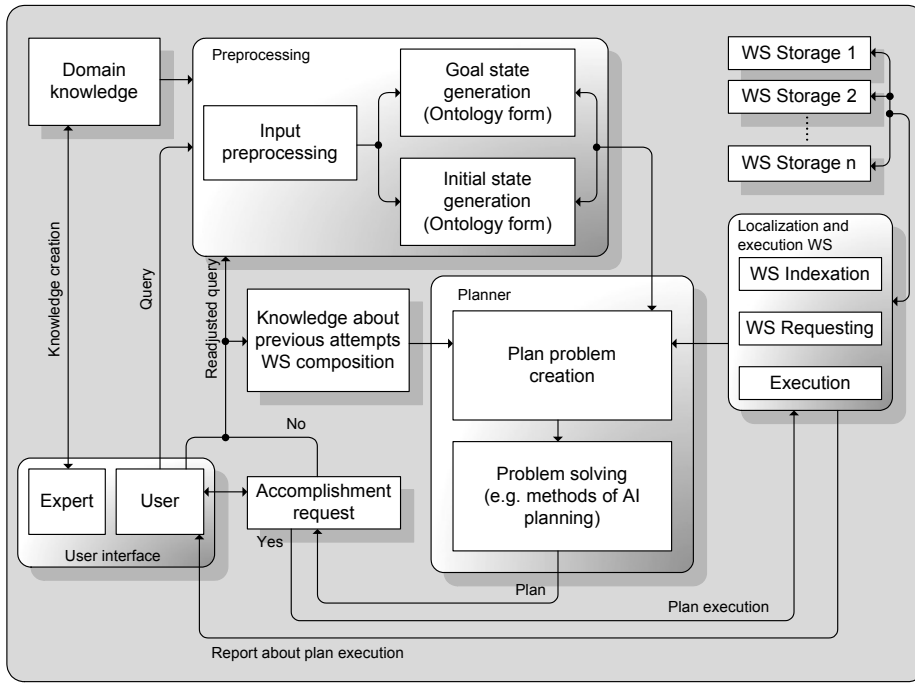


Fig. 9. System proposal for web services composition

6. Actual problems and web services requisites

One of the problems of actual systems for automatic web services composition is their control complicacy. For result and full process of planning is required have correct described initial and goal state. Next is convenient note, that full automatic system creation is complicated and the more we will try make system more automatic, thereby here emerges increasingly problems. E.g. the system will be specified increasingly only to fixed field of solving problems, or the system will not offer suitable results and etc. Therefore it will be convenient implement into system also some possibility of interaction with user, in order to extend this field in case of need. E.g. OWLS-Xplan allows to user dynamically interfere with plan during planning. These interferences but system don't save and in the case of repeated creation the same plan with the same conditions will be work alike. Here it will be convenient try make some possibility of system learning. It will be also convenient e.g. composition process and also result represent by graph, where nodes could be states and edges could be actions (web services). This graphics representation could bring to system bigger credibility from users.

Ideal system for web services composition should contain following part:

- should provide easy possibility ontology addition for planning problem.
- possibility for web services searching and indexing, eventually easy possibility web services registration by user.
- following ontologies should user easy describe initial and goal state. In ideal case would initial and goal state obtaining at the most automatically (e.g. by some domain would obtain a part of necessary information from user profile).
- should contain executive tool for web services composition. Artificial intelligence methods appear as most suitable.
- should enable dynamically interfere with planning and change plan to user. How was remember above, suitable representation is by graph. In the case of necessity user should have view how progress actual planning and could interfere with its following own reflection.
- system should show some intelligence. Intelligence might inhere e.g. in saving some knowledge obtained from user interference during planning.

Acknowledgement

The work presented in this paper was supported by the following projects: the Slovak Research and Development Agency under the contract Nr. APVV-0391-06; the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant Nr. 1/4074/07.

References

[1] WS: Web Services Glossary, http://www.w3.org/TR/wsa-reqs/, 2002

[2] RAO, J. – SU, X.: A Survey of Automated Web Service Composition Methods. In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition, SWSWPC 2004, San Diego, California, USA, 2004.

[3] NILSSON, Nils J. – FIKES, Richard E.: STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving, Artificial Intelligence, 2(3):189-208, 1971.

[4] WU, Dan – SIRIN, Evren –HENDLER, James – NAU, Dana – PARSIA, Bijan: Automatic Web services composition using SHOP2, In Workshop on Planning for Web Services, 2003

[5] PEER, Joachim: Web Service Composition as AI Planning - A Survey, Dissertation, University of St. Gallen, Switzerland, 2005.

[6] SIRIN, Evren – BIJAN, Parsia – WU, Dan - HENDLER, James: HTN planning for web service composition using SHOP2, Journal of Web Semantics 1 (4), pp.377-396, 2004

[7] KLUSCH, M. - GERBER, A. - SCHMIDT, M.: Semantic Web Service Composition Planning with OWLS-Xplan.1st Intl. AAAI Fall Symposium on Agents and the Semantic Web, Arlington VA, USA, 2005

[8] OH, S. - LEE, D. - KUMARA, S. R.: A comparative illustration of AI planning based web services composition. ACM SIGecom Exch. 5, 5 Jan., 2006

[9] BLUM, Avrim L. – FURST, Merrick L.: Fast planning through planning graph analysis, In Artificial Intelligence journal volume 90, 1636 -1642, 1997

[10] SIRIN, Evren – HENDLER, James – BIJAN, Parsia: Semi-automatic composition of web services using semantic descriptions in Web services: Modeling, Architecture and Infrastructure workshop in ICEIS 2003, Angers, France, April 2003

[11] REITER, Ray: On knowledge-based programming with sensing in the situation calculus. ACM Trans. Comput. Logic 2, 4 (Oct. 2001), 433-457, 2001

Secure Process-oriented Infrastructure for Networked Enterprises

Karol Furdík

InterSoft, a.s., Floriánska 19, 040 01 Košice, Slovakia
karol.furdik@intersoft.sk

Abstract. The paper describes the FP7 ICT EU project SPIKE, aiming at the development of a software platform for secure, flexible, and project-oriented collaboration of business organisations within a temporary alliance. The project objectives, overall vision of the solution, and related research approaches are presented as a background for designing the system architecture. Functional components identified for the SPIKE platform are described together with the respective technology foundations. Focus is especially given to the integration of service-oriented architecture, business process modelling, and semantic technologies. The paper presents the system prototype as it was designed for the first trial of pilot applications.

Keywords: Business alliances, networked enterprises, interoperability of services, ontology development, semantic business process modelling.

1 Introduction

A flexible and effective inter-organisational collaboration is nowadays widely accepted as an important success factor in the globalised knowledge economy. In the context of progressive technologies such as service-oriented architecture (SOA), semantic web services, and business process modelling, it was also addressed by the European Commission in its 7th Framework Programme, Challenge 1: "Pervasive and Trustworthy Network and Service Infrastructures" [5].

The *Secure Process-oriented Integrative Service Infrastructure for Networked Enterprises* (SPIKE, www.spike-project.eu) EU ICT project No. FP7-ICT-217098 is one of the responses on this challenge, aiming at the design and implementation of a system for enterprises of all sizes to enable a creation and maintenance of project-oriented short-term business alliances of collaborating enterprises. The project, coordinated by the University of Regensburg, Germany, started in January 2008 and is planned to run for 3 years. The project consortium consists of eight partners from five European countries, commercial enterprises as well as universities. The Slovak project partners, i.e. Technical university of Košice and InterSoft, a.s., are responsible mainly for tasks related to the system design and implementation.

The main goal of the SPIKE project is to research and implement a system that will bring flexibility to the collaboration between networked enterprises. Particular objectives of the project were specified on organisational as well as on scientific and

technological levels. The organisational objectives are focused on a simplification of business collaboration through dynamically created and pre-defined business processes and workflows. The solution should provide the service interoperability and should enable the outsourcing of parts of the process value chain to business partners in short-term business alliances.

The scientific and technology objectives include research, development, implementation, and validation of the components for semantically enhanced business process management environment. Namely, it covers components such as semantic service bus, semantic business process modelling engine, information flow control between the alliance members, security infrastructure, storage repositories for processes and ontologies, as well as a portal-based interface providing user-friendly administration, maintenance, and utilisation of alliances.

1.1 SPIKE Approach Towards the Networked Enterprises

According to defined project objectives, the envisioned functionality of the SPIKE system is to provide a technical support for collaboration of business partners aiming to create a temporary business alliance. Three phases of the alliance life cycle [6], i.e. the *setting-up* a new business alliance upon a specified project, *running* the alliance according to a defined workflow, and *closing down* the collaborative project, are supported with respect to the three levels of collaboration:

- *Collaborative processes* that enable to produce physical or intangible artefacts and are modelled by means of complex workflow patterns;
- *Sharing services*, where the alliance partners can offer their services in the scope of a given business process. The offered services can be retrieved, negotiated, contracted, and finally used by the alliance members according to the conditions specified by the service contract;
- *Identity federation*, enabling and mediating the access of an alliance partner to the internal resources or services of other partners.

The SPIKE project adopts the approach that aims to support all the mentioned collaboration levels. High-level picture of this approach is schematically depicted in Fig. 1. Business organisations, presented in the top bar and labelled as “SPIKE Alliance”, may decide to form a new alliance that is focused on the production of a concrete artefact. The first step is to define a collaborative value chain, which determines a target and particular steps of the short-time alliance. The value chain, depicted in the middle bar of Fig. 1, is then expressed as a business process and is formally modelled by the BPMN notation [16]. To enable the interoperability of possibly heterogeneous capabilities, knowledge, data, and services of each of the alliance partners, a common and shareable knowledge model is used for enriching the defined business process with semantic information [9]. The resources and services of participating organisations may then be mediated and integrated according to known and formalised meaning, represented by the shareable ontology.

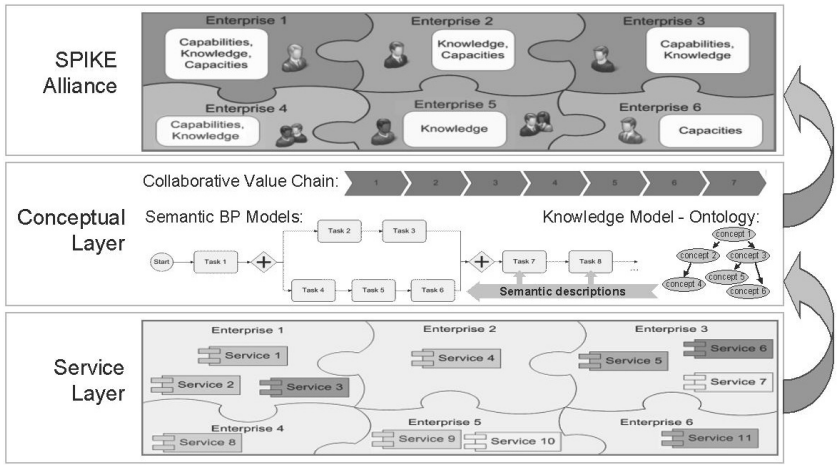


Fig. 1. Basic schema of the SPIKE approach towards the networked enterprises.

Particular tasks in the process model are then grounded to the executable services provided by the alliance partners (see bottom bar of the Fig. 1, labelled as "Service Layer"). It allows sharing and using the services in the context of a defined process model by authorised organisations. Assuming that some of the services may access or manipulate the internal infrastructure of an alliance partner, identities and credentials necessary for consuming such a service are distributed to the authorised users in a secure way. The alliance can then operate according to a dynamic process model, which, if needed, may be modified and adapted in the run time.

1.2 Related Research

The SPIKE approach is enabled and will be supported by several advanced technological and conceptual achievements, namely by the principles of SOA, concepts of web services and semantic web [2], security and identity management, knowledge representation, and semantic process modelling. A wide variety of solutions, tools, and approaches exist nowadays in these fields as outcomes of research projects and even as successful commercial products. To reduce necessary implementation efforts and to focus on the innovative design, as well as due to compatibility reasons, the SPIKE project is aiming to reuse existing and available high-quality solutions, mostly taken from the outcomes of the related EU projects. Namely, the following EU research projects were identified as these of particular interest for SPIKE:

- *STASIS* (FP6-034980, www.stasis-project.net): provides a platform for semantically enhanced eEconomy services, helps to achieve the semantic interoperability and data mediation;
- *TrustCom* (FP6-001945, www.eu-trustcom.com): provides a framework for virtual organisations, namely the trust, security, and contract management;

- *SeCSE* (FP6-511680, www.secse-project.eu): supports service-centric applications, especially the specification, discovery, design, and management of services;
- *OPUCE* (FP6-034101, www.opuce.tid.es): provides an unified and open service environment, which can serve as an infrastructure for collaborative and dynamic loosely coupled services;
- *SUPER* (FP6-026850, www.ip-super.org): provides a modular architecture for semantic modelling of business processes.

These projects treat special aspects, which SPIKE project aims to extend and integrate into a comprehensive, ready-to-use solution for building business alliances.

2 Architecture, Functional Components and Data Structures

The design of the architecture proposed for the SPIKE platform was accomplished in line with the methodology of Rozanski and Woods [13], [6]. The viewpoints, perspectives, and stakeholders were identified for the overall system. Then, based on the description of required functionality, which was provided by the user partners of the project [15], the system scope and context were specified. The scope of the SPIKE system, which roughly corresponds with the functional viewpoint, is determined by the envisioned functionality of the system as whole.

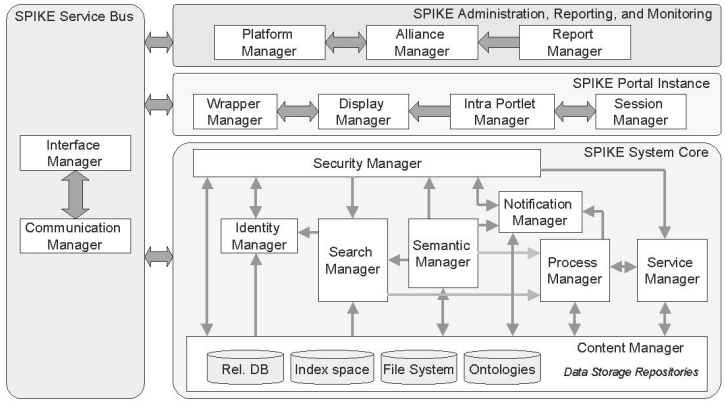


Fig. 2. Architecture and functional components of the SPIKE system.

The highest level of the functional system architecture, designed in line with the SOA principles as highly modular and extensible, is schematically depicted in Fig. 2. It consists of four main functional subsystems:

- The *SPIKE System Core* (SSC) is a back-end that provides functions for processing all the system data. It includes components for data storage and retrieval, security and identity management, maintenance of processes, workflows, and services, as well as components for semantic metadata description and ontology maintenance.

- The *SPIKE Portal Instance* (SPI) is a graphical user interface and acts as a front-end to the SSC. It combines multiple web applications, so-called portlets, into one single portal web page. The SPI provides components for handling user sessions and events generated by SSC components, visualisation of the services connected to the platform via SSB, and integration of external legacy applications that do not offer a service-oriented interface.
- The *SPIKE Administration, Reporting, and Monitoring* (SAMR) subsystem is a toolkit for system maintenance and day-to-day operation, including administration and management of business alliances, reporting facilities, and monitoring of the whole SPIKE platform.
- The *SPIKE Service Bus* (SSB) enables internal communication between SSC, SPI and SAMR as well as communication with external entities.

Each of the subsystems consists of several *managers* – system components that provide autonomous and elementary functionality. The design of managers was accomplished according to the methodology of [3], namely trying to balance the coupling vs. cohesion and sufficiency vs. completeness metrics.

All the subsystems in the platform were decomposed into 17 different managers, which have been broken down into 48 internal modules [8], [1]. The managers and their modules were described in detail; the specification consists of the context of a manager, supported use cases, and a structure of the manager's modules including APIs, dependencies, and mutual interactions.

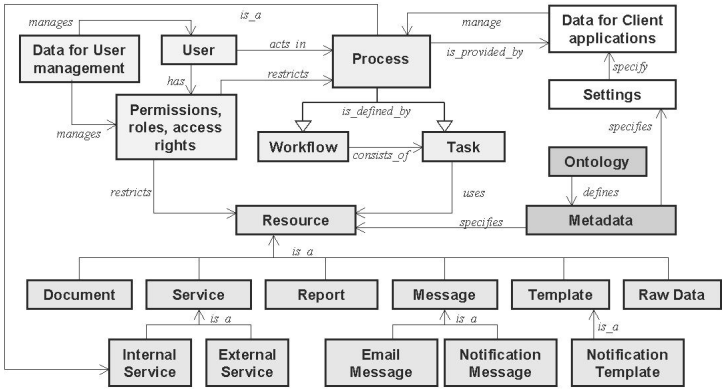


Fig. 3. The structure of basic data elements, designed for the SPIKE system.

The Fig. 3 presents the structure of main data elements, as it was identified in the information view of the SPIKE system architecture. The *Process*, *Workflow*, and *Task* elements are basic building blocks for modelling an alliance of collaborative business processes. The *Task* element, representing particular workflow actions, is further specified by parameters as inputs, transformations, and outputs. These parameters, consumed and produced by a task in a workflow, are represented by a set of sub-types of the generic and abstract *Resource* data element. It defines a set of common properties that are inherited by all the child data elements, in particular by the

resource types as *Document*, *Service*, *Report*, *Message*, etc. Properties of these information resources are provided as semantic metadata, defined in the ontology schema. This solution enables to combine the standardised business process modelling with semantic descriptions created according to the Semantic Web principles [2].

Semantic technologies are employed in the SPIKE platform to enable the interoperability between services and information resources originating from potentially heterogeneous environments of the alliance partners. The *ontology* serves as a common semantic knowledge base and a conceptual model of the given domain. It provides a shareable vocabulary of terms that can be used for *metadata* specification of used data elements (cf. Fig. 3). The tasks and services, i.e. the elements of a business process modelled by a workflow, can also be semantically enhanced, described by the ontology concepts. It gives an opportunity to manipulate with the workflow data according to their meaning; namely, it enables data mediation, reasoning, and retrieval of the semantically described (i.e. annotated) elements and consequently ensures the desired service interoperability.

The structure of ontologies designed for the SPIKE platform is based on the identified data elements and information resources, as depicted in Fig. 3. It includes a separate ontology for *business processes*, *services*, and *artefacts*. Four additional ontologies were specified for modelling the domain- and system-specific information, namely the *core*, *domain*, *system*, and *user* ontologies. The conceptual model of adopted semantic framework (i.e. the WSMO, see in next section) together with the existing ontology resources reused and adapted from external ontology libraries and standards (e.g. Dublin Core, ontologies provided by the FP6 project SUPER, etc.) were used as resources in the ontology creation process. However, the *requirement-driven approach* [10] was employed as the main methodology for the ontology design. Based on this approach, the requirements formulated by user partners of the project were collected in a systematic way, formalised, and implemented in the WSML ontology language. More details on the design, development, and life-cycle management of the SPIKE ontologies can be found in [7].

3 Technology Employed for the System Development

The SPIKE subsystems can be, from the implementation point of view, divided into three packages that covers design tools, user interface, and core system. Each of these packages employs a specific technological framework; however, integration of these frameworks is ensured by the employed SOA and web service interfaces.

The package of *design tools* corresponds to the functionality of SAMR sub-system. Technologically, it is based on the *Eclipse IDE* platform (www.eclipse.org). It consists of tools for BPMN modelling (*WSMO Studio*, *BPMO designer*), including facilities to export an abstract BPMN model into executable BPEL processes [16]. In addition, the package contains tools for designing the interfaces for the tasks performed by a human actor (so-called human tasks), implemented using the XForms (*Visual XForms Designer*). Administration tools also include the toolkit for designing the ontologies and semantic mediator rules, as well as the tools for semantic

annotation of services using the SA-WSDL annotations (*WSMO Studio*, www.wsmostudio.org).

The package of *user interface* functionally corresponds to the SPI sub-system. It is designed as a set of portlets integrated in the standard *JSR 168 Portlet container*. The user interface provided by the *Intalio Tempo* framework (www.intalio.org) for the BPEL4People processes is employed in the portlet component responsible for displaying the human tasks.

The *system core* package covers the functionality of both SSB and SSC sub-systems. It is designed as the *JSR 208 Java Business Integration* (JBI) compliant enterprise service bus (ESB). The *Apache ServiceMix ESB* was selected as a suitable technology for the JBI container implementation. The BPEL execution engine, provided by the Process Manager (cf. Fig. 2), is designed as a JBI component extended on the semantic binding feature. The standard *Normalized Message Router* (NMR) of the JBI serves as the main communication channel for delivering messages between the SSB components.

The semantic functionality, namely the data mediation, retrieval, composition, and orchestration of services, is supported by the software packages developed for the *WSMO Lite* semantic framework [14]. In particular, the *wsmo4j* package provides an access to the in-memory object model of the WSML ontology elements. The *wsml2reasoner* API is used as the underlying inference engine (i.e. a reasoner). Implementation of the ontology repository is based on the *Ontology Representation and Data Integration* (ORDI, http://www.ontotext.com/ordi/) framework.

The SPIKE security infrastructure, identified as one of the most important and crucial factors for the business collaboration, was designed in terms of attribute/role management. Security features are implemented in the Security Manager, which can be seen as a horizontal layer that influences most of the platform components and managers. It supports the authentication, workflow and service access control, and auditing functionality. The authentication is solved by a hybrid mechanism of *SASL* [12] and *GSS-API* [11], integrated with the WS-Security protocol for secure web services. The *PERMIS* infrastructure [4] was selected to handle the authorisation, namely the facilities of managing users' privileges and authorisation policies. The auditing employs the non-repudiation approach, which provides the compulsory certified tracing by means of confirmation services and timestamps. In future, this solution enables to extend the security infrastructure by adding the digital signatures into the service communication.

4 Conclusions

The SPIKE platform, presented in this paper, employs the principles of SOA, business process modelling, and semantic technologies, supported by a strong security infrastructure. It is aimed at creation and maintenance of temporary and project-oriented business alliances, enabling flexible collaboration of involved enterprises.

The paper gives an overview of the architecture design, functionality, and technology adopted for the platform implementation, as it was accomplished in the first half of the FP7 ICT EU project SPIKE. The implementation of the designed

platform is currently ongoing and will result in the first prototype, which should be available in September 2009. The prototype will then be tested within the first trial on three pilot applications in business organisations from Austria and Finland. More information on the SPIKE project can be found at www.spike-project.eu.

Acknowledgments. The SPIKE project is co-funded by the EC within the contract No. 217098. The work presented in the paper was also supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic within the 1/4074/07 Project "Methods for annotation, search, creation, and accessing knowledge employing metadata for semantic description of knowledge".

References

1. Bednar, P. et al: D5.1: Spec. of Components for the Service Bus Sub-system. Technical report of the SPIKE project, FP7-ICT-217098. Technical University of Kosice (2009)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. In: Scientific American (2001)
3. Booch, G. et al: Object-Oriented Analysis and Design with Applications, Third Edition, Addison Wesley Professional (2007)
4. Chadwick, D. W., Otenko, A., Ball, E.: Role-Based Access Control with 470 X.509 Attribute Certificates, IEEE Internet Computing, vol. 7, No. 2, 62--69 (2003)
5. CORDIS ICT Challenge 1: Pervasive and Trusted Network and Service Infrastructures. http://cordis.europa.eu/fp7/ict/programme/challenge1_en.html. Accessed: 09/08/2009
6. Furdik, K, Mach, M, Sabol, T.: Architecture of a system supporting business alliances. In: Proceedings of WIKT 2008, pp. 53--57, STU, Bratislava, Slovakia (2009)
7. Furdik, K, Mach, M, Sabol, T.: Towards semantic modelling of business processes for networked enterprises. In: Di Noia, T., Buccafurri, F. (eds.) E-Commerce and Web Technologies. 10th International Conference, EC-Web 2009, Linz, Austria, September 2009. Proceedings. LNCS, vol. 5692, pp. 96--107. Springer, Heidelberg (2009)
8. Gmelch, O. et al: D4.1: Specification of components for Portal System. Technical report of the SPIKE project, FP7-ICT-217098. University of Regensburg (2009)
9. Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management. In: Proceedings of the IEEE ICEBE 2005, October 18-20, Beijing, China, pp. 535--540 (2005)
10. Klischewski, R., Ukena, S.: Designing Semantic e-Government Services Driven by user Requirements. In: Electronic Government, EGOV 07 conference. Proc. of ongoing research, project contributions and workshops, pp. 133--140, Trauner Verlag, Linz, Austria (2007)
11. Linn, J.: Generic Security Service Application Program Interface, Version 2, Update 1. RSA Laboratories (2000)
12. Melnikov, A., Zeilenga, K.: Simple Authentication and Security Layer (SASL), Isode Limited, OpenLDAP Foundation (2006)
13. Rozanski, N., Woods, E.: Software Systems Architecture. Working with Stakeholders Using Viewpoints and Perspectives. Addison Wesley (2005)
14. Vitvar, T., Kopecky, J., Fensel, D.: WSMO-Lite: Lightweight Semantic Descriptions for Services on the Web. WSMO Deliverable D11, Ver.0.2. DERI (2008)
15. Vogler, H. et al: D2.2: User requirements analysis and dev./test recommendations. Technical report of the SPIKE project, FP7-ICT-217098. IT Inkubator Ostbayern GmbH (2008)
16. White, S. A.: Using BPMN to Model a BPEL Process. IBM Corporation (2005)

Author index

František Babič, Ján Paralič, Andreas Rauber
Editors

Proceedings

9th International Student Workshop WDA 2008

1st edition, 70 copies, Published by Centre for Information Technologies, Faculty of Electrical Engineering and Informatics, Technical University in Košice, Slovakia

Printed by EQUILIBRIA, s.r.o.

2009

ISBN
EAN