



Briefing paper: Value of software agents in digital preservation

Ver 1.0

Dissemination Level: Public

Lead Editor: NAE

2010-08-10

SEVENTH FRAMEWORK PROGRAMME THEME

ICT -1-4.1 Digital libraries and technology-enhanced learning

Project full title: PReservation Organizations using Tools in AGent Environments

Grant agreement no.: 216746



Table of Contents

1. Introduction.....	4
2. The PROTAGE prototypes	5
2.1. First PROTAGE prototype.....	5
2.2. Second and third PROTAGE prototypes	6
3. Potential of software agents in digital preservation	7
3.1. Agents in pre-ingest	7
3.2. Agents offering digital preservation guidance.....	8
3.3. Monitoring agents	9
4. Conclusions	10

1. Introduction

This briefing paper discusses the potential use of software agents in digital preservation. The paper builds upon the experiences drawn from the development of the PROTAGE prototypes and takes mainly the view of a National Archives.

The paper includes:

- a short description of the PROTAGE prototypes;
- a discussion on the potential of the prototypes for the intended stakeholder groups;
- conclusions of the discussion.

2. The PROTAGE prototypes

During the project, the PROTAGE team developed three prototypes which were aimed at different areas of the digital preservation task of stakeholders. Therefore, the prototypes took conceptually different approaches towards applying software agents to digital preservation. The prototypes and their differences are further explained in the following sections.

When reading through the following sections, please take into account that this paper is not intended to give a detailed technical overview of the software agents developed in PROTAGE. If you would like to learn more about the technical side of the agents, please refer to the technical documentation of the project.

2.1. First PROTAGE prototype

During the first year of the project, the PROTAGE team organised a set of need-finding events mainly oriented towards the main actors in digital preservation– currently memory institutions and governmental agencies. At these need-finding events, the main option was that the PROTAGE prototype should help governmental agencies in meeting pre-ingest and transfer requirements specified by memory institutions.

This was also the starting point for the first PROTAGE prototype, which was based on three use scenarios:

- Checking SIP requirements;
- Creating an overview of files to transfer;
- Creating SIPs ¹

It was planned that users should be able to use the prototype to get access to SIP structures as XML Schemas, check their metadata against the SIP requirements as well as compare file formats to those allowed by the memory institution for transfer. The PROTAGE environment also intended to include a few web services which could be used to create missing metadata, check files for viruses and convert files into archival file formats.

The intention of the PROTAGE project was to investigate the value of agents especially for finding the suitable SIP format for an agency, determining and applying the correct software to create the SIPs.

The first prototype was developed to be a web-based environment accessible from an ordinary web browser. It supported one SIP format as an XML schema and a limited number of tools as web services. It also has to be mentioned that due to the technical complexity of the task at hand, especially the complexity of metadata requirements from national archives, the prototype did not fully succeed in meeting all the intended functionality.

¹ For a definition of the term SIP (Submission Information Package), see the standard ISO 14721 (Reference Model For an Open Archival Information System)

2.2. Second and third PROTAGE prototypes

After testing the first prototype and evaluating the feedback, the project decided to change the direction and no longer concentrate on the pre-ingest actions, but rather develop a digital preservation tool which could be used in whatever digital preservation tasks. For this purpose, the project took up a workflow-based approach where users are enabled to combine digital preservation tools that are known to the software agents into digital preservation workflows, called Action Plans. The main intention was to apply this approach to the monitoring of a user's digital assets. During the second year of the project, a new set of application scenarios was also developed to reflect this. In addition, the second and third prototypes were more oriented towards individual home users and were intended to help the users in determining the long-term risks associated with their digital files and offer best-practice solutions for dealing with those risks.

The second and third PROTAGE prototypes were developed as a downloadable and locally installable client application which:

- Enables a user to define his/her collections, user profile and friends (i.e., the social network of the user for digital preservation);
- Enables a user to define his/her digital preservation related knowledge in the form of workflows, the descriptions of preservation tools and web services;
- Allows a user to search for digital preservation related knowledge (digital preservation workflows, tools and web services) by exploiting his/her social network;
- Orders the search results in a recommendation order by using the trust levels of the recommendations and users available in the user's social network;
- Executes workflows with the help of public web services or locally downloaded digital preservation tools;
- Monitors the user's file formats and recommends migration actions if necessary.

In addition to the locally installed client application, the PROTAGE environment included so-called Access Points which can be viewed as digital preservation experts who gather information about preservation tools, combine preservation workflows and make those available to ordinary PROTAGE users. The PROTAGE Access Points are also enabled to exchange information about the best practices of digital preservation, similar to the PROTAGE Client Applications.

The third PROTAGE prototype was an incremental improvement of the second prototype mainly addressing the integration aspects, such as the coordination with the EU project PLANETS (ends during 2010) and other preservation environments. The testing was this time based on the third prototype and partly made in an iterative way. It means that the test results were continuously fed back to the development team in order to finalise the PROTAGE prototype.

3. Potential of software agents in digital preservation

As described in the previous chapter, the first prototype, in relation to the other two prototypes, was conceptually different. The test phases and feedback also indicated different opinions about the potential of software agents. Thus, it is reasonable to take the discussion following the different parts of digital preservation:

- the use of agents for agencies and companies in pre-ingest activities (first prototype);
- the use of agents in a digital preservation guidance function (second and third prototype);
- the use of agents in a monitoring function (second and third prototype).

3.1. Agents in pre-ingest

From the point of view of a public agency or a private company, the essential characteristic of pre-ingest activities is that there is a clearly defined long-term repository, whether in-house or external, which takes care of the user's digital assets after pre-ingest and the execution of transfer activities. Additionally, the timeframe and requirements for transfers (and thus also pre-ingest actions) are usually clearly defined by a transfer policy which describes when and how the user has to transfer his/her digital assets to the repository. It is also worth mentioning that the stakeholders making use of a long-term repository usually have a rather large amount of digital data to be transferred.

The software agents in the first PROTAGE prototype were modelled to take those characteristics into account. They were meant to gather information about the user's location and affiliation, thereafter decide on the legislative environment for the given situation and the repository to use, download SIP requirements and finally fulfil all necessary tasks to assure the necessary quality of the data to be transferred.

After getting the results of the test and evaluation of the first prototype and looking deeper into the applied scenarios, the project was able to identify multiple possible arguments against the application of software agents in those scenarios. Firstly, the selection of a suitable long-term repository is usually a straightforward task. As there are appraisal and consultation routines already in place in most countries, the public agencies are rather well aware about the transfer conditions of the repositories even without the help of software agents. Of course, for private companies this is not always the case. But, for them the repository selection process does also include pricing and contract negotiations. While software agents could potentially be used for such negotiation purposes, the development of these agents is not a straightforward task and therefore the agent-based search possibility probably offers little benefits over a simple Google or Yellow Pages search.

The second reason is the stability of the SIP requirements. Usually the repositories do not change their transfer requirements too often and thus agencies and companies with an established contract to a repository can develop the needed pre-ingest functionality locally without using agent technology and continuously use it without major changes for multiple years. Simply put – using agents is more suited in dynamic environments and the efforts put into their development can show non-trivial benefits when changes in procedures and requirements occur more often. Taking into account also the technical complexity of applying agents in a controlled legislative and technical environment (for example, most transfers occur from an EDRM system or some database. Thus agents used in a corporative environment have to be able to “read” information from those systems) it seems more effective for the agencies or the repositories themselves to develop purpose-built tools to meet the exact requirements, instead of using a combination of web based services usually

developed by other stakeholders and probably not always fully matching the requirements of the repository. In particular, the complexity of metadata structures and respective metadata requirements of repositories pose a challenge which is not easily solvable by software agents.

Finally, during the test phase of the first prototype the sensitivity of corporate data was also mentioned as an important reason for not using agent technology, but employing locally installed solutions which are under the control of the agency and thus more trusted.

As a positive argument, the application of software agents to identify tools which can be used in the pre-ingest process additionally to the purpose-built tools seemed promising as of the first PROTAGE prototype. As an example, agents could be used to find and exploit migration tools or web services for less common file formats when neither the agency nor the repository have sufficient knowledge for developing good quality tools themselves.

To conclude, the first PROTAGE prototype did not succeed to demonstrate the potential of agents with the intended stakeholder group and scenarios mostly because of the technical complexity of the corporate systems and the transfer rules, but also because of legislative and sensitivity issues.

3.2. Agents offering digital preservation guidance

Drawing on the outcomes of the development and test of the first PROTAGE prototype, it was decided that the following development should concentrate more on the application of agents for helping users determine the best possible workflow for dealing with digital preservation related risks and select the appropriate tools for the workflow steps. Currently, users who have identified a digital preservation risk use different information sources to gain necessary knowledge to deal with it: "Google search" can probably be regarded as the most used method. But e-mails and phone calls to experts in memory institutions are also considered to be useful.

Therefore, the second and third PROTAGE prototypes include an agent-based guidance functionality which takes a step beyond usual Google searches: The user is enabled to analyse his/her digital assets which are available as simple computer files using packaged standard components (like Droid or JHove). Thereafter, the user is able to initiate keyword-based searches which through agent communication try to find suitable preservation tools by keywords, but also matching the user and collection characteristics. To put things simple: PROTAGE enables a user to find hits for a keyword from his/her trusted friends who have similar user needs and similar collections.

Compared to the agents discussed in the previous chapter such a guidance agent (called SearcherAgent in PROTAGE) is simpler, mainly because it does not try to deal with an explicit digital preservation scenario but allows searches for any digital preservation workflow and/or tool by using an intelligent, trust-based keyword search.

From the point of view of a digital preservation expert such an agent has many benefits. First, it allows memory institutions (or in general digital repositories) to put information about singular digital preservation tasks into the PROTAGE environment. It can then be reused for both their clients and other PROTAGE users in the future. Also, the memory institutions themselves have a quick and effective way of acquiring bits of missing knowledge about available digital preservation tools and methods. Therefore, it limits the number of manual inquiries made to the repositories' digital preservation experts.

Second, the quality of the trust-based keyword search is obviously superior to Google or similar search engines by delivering less results of higher quality due to the fact that the ratings of users are used to filter and recommend the results in a personalized manner. Therefore, the guidance agent limits the time needed for going through the results and figuring out which one is really the best solution.

Third, the agent in PROTAGE also benefits from a feedback element – users who have executed some tools that were found with the help of PROTAGE are able to give their opinion about the quality of the tools. As with the agent-guided keyword search, the feedback is also taken into account based on the individual context of a user following the argumentation that negative (or positive) feedback from one user does not necessarily imply to all user groups and all kinds of digital data.

From the negative perspective, it is visible that the number of participating memory institutions has to be rather large in order to include a good set of recommendations. Also the users have to be quite active as the methods for preferring some recommendations over others are rather quantitative (a good recommendation has to get multiple positive ratings to prevail over others). Additionally, a best practice advice does not always come in the form of tools to be used, but sometimes as some articles, blog posts or a list of recommendations on how often to repeat your preservation actions. Currently this “softer” knowledge is not present in PROTAGE and therefore the amount of knowledge in general is limited.

3.3. Monitoring agents

Another negative aspect of the guidance agent is that the user has to know the right keyword. In other words, in order to get the right answer, the user has to know the right question first. Looking at the current situation, this means that the guidance agent described above is only useful for digital preservation specialists who are well aware about the risks in digital preservation and are seeking for answers to those.

To deal with this problem, the PROTAGE second prototype includes a monitoring agent called the “MigrationSupportAgent”. This agent takes a novel approach to determining migration needs by collecting statistics about file formats used in a user’s network. Based on the quantity of available file formats the user is intended to decide whether a file format is too less used or not and based on the decision the user can then start looking for advice to migrate the respective files into a new, more common, file format.

The potential of such a monitoring agent is rather clear and it is obvious that the given quantitative effect on selecting an archival file format is well covered by software agents. It is also rather simple to automate this functionality in such a way that agents automatically detects file formats at risk (for example, when less than 5% of users in the network use a file format), also automation of the search for an appropriate migration tool is technically rather simple.

From the negative perspective, the MigrationSupportAgent requires a rather large user community to function correctly and to deliver good argumentation for the decision.

4. Conclusions

From the discussion above, several preliminary conclusions can be drawn as follows:

- Software agents justify themselves most in the situations where users are rather spread and do not have well known and strict requirements (especially metadata requirements) to follow. Therefore, it seems most reasonable to use agents in situations where new knowledge is needed for taking some actions (for example, maintaining private file collections, doing digital preservation research) and less so when there is already an established digital preservation routine available (for example, transfers to national archives).
- The application of automated digital preservation tools on corporate information systems can be regarded as being technically complex. Therefore, the PROTAGE project did not succeed in showing the benefits of software agents over purpose-built specific preservation tools or functionality.
- The application of software agents on larger workflows like pre-ingest seems not reasonable, while the “agentification” of singular preservation tasks regardless of their position in a larger workflow seems to be more promising.
- The potential of software agents is clearly visible in a guidance function and most in a monitoring function, as they are able to react on the changes in the environment, and are also able to take into account the different contexts of different users.
- Software agents are highly dependent on the number of agents in the environment, because most of the potential lies in applying quantitative mechanisms and feedback.
- The agent environment is also highly dependent on the knowledge available in it. Therefore, the take-up of PROTAGE by expert users who contribute to the environment is crucial.

As an additional comment, it has to be said that the scope of the PROTAGE project was limited in the sense that it did not include research in semantic tools and applying those together with software agents. However, the existing methods and tools which can be used to semantically analyse users’ collections would potentially help take a next step, both when it comes to acquiring more information into the agent network from external sources (e.g., digital preservation forums, blogs and wikis) and to enabling better results in “keyword guidance” for users who are not well aware about risks in digital preservation. Additionally, using semantic tools along with agents could probably make it possible to apply software agents in automated metadata creation and reuse.