DPÜ Topics for 2013

1: Improve quality and coverage of characterisation results using FITS (2x2)

Context

Characterisation tools don't always agree on either identification (which formats the objects are) nor other elements (such as validity) of files, and hence produce conflicting results when combined (such as in the way *FITS* does). Finally, some tools are known to be rather useless or unstable on certain file types, but FITS calls them in the default configuration. All this potentially creates a messy data set, in particular if large-scale analysis is the goal. Conflicting values are hard to reason on, so improving on this is a critical step towards preservation analysis, planning, and actions.

Goal

Analyse conflicted data and derive priority and mapping heuristics to configure FITS in a way that re-running it reduces the number of conflicts. Re-run FITS on the data set, analyse the resulting performance, files, compare the two sets, evaluate the improvements quantitatively, and discuss the improvements, limitations, challenges.

Input

Please use the govdocs1 dataset which you can download from http://digitalcorpora.org/corpora/files. For this exercise we encourage you to use the 10 subsets of the corpus which you can download from http://digitalcorpora.org/corp/nps/files/govdocs1/zipfiles/ (thread0.zip ... thread9.zip). Each subset contains 1000 files, randomly selected from the corpus, making up 10.000 files in total.

Concept

Describe your preliminary analysis of common conflicts, outline a statistical approach to verifying that the additional rules do not destroy valuable knowledge, and include a short outline of the rules available in FITS and how they can be used to reduce conflicts. Also specify the approach you are using to evaluate your improvements. For evaluation of the exercise we will apply your rules to a different set and evaluate the improvements independently.

Results

To submit

- FITS configuration
- Complete set of FITS files (created with your configuration) from the govdocs1 files
- Statistical analysis results including Type I/II errors.
- Report discussing improvements made, limitations, and challenges

To present

- A few typical conflicts and how they have been resolved
- Statistics about improvements, error rates. What about the long tail of rarely occurring conflicts that are not captured by the specific rules? How could they be tackled?

Contact For questions, contact Hannes Kulovits (kulovits@ifs.tuwien.ac.at)

2: Reduce conflicts using post-processing rules in c3po (2x2)

Context

Different characterisation tools don’t always agree on either identification (which formats the objects are) nor other properties (such as validity and well-formedness) and hence produce conflicting results. This creates a messy data set which is hard to interpret, especially when large-scale analysis is the goal. Conflicting values are hard to reason on, so improving on this is a critical step towards preservation analysis, planning, and actions.

Goal

Develop and apply post-processing methods and tools to FITS characterisation files within c3po to reduce the problem of “conflicting values” and significantly improve data quality. Evaluate these for improvement and validate in how far they do not cause wrong statistics (Type I / Type II errors, etc.). Discuss the limitations of this approach, and think how the long tail of conflicts could be addressed.

Input

Please use the govdocs1 dataset which you can download from http://digitalcorpora.org/corpora/files. For this exercise we encourage you to use the 10 subsets of the entire corpus which you can download from http://digitalcorpora.org/corp/nps/files/govdocs1/zipfiles/ (thread0.zip ... thread9.zip). Each subset contains 1.000 files, randomly selected from the corpus, making up 10.000 files in total.
Fork from the c3po repository which can be found at: https://github.com/peshkira/c3po

Concept

Discussion of the design of the c3po modules that are affected, and in which way they can be adapted. Discussion of the key issues and how they will be tackled. Outline a statistical approach to verifying that the post-processing does not destroy valuable knowledge, and include a short outline of the approach you are using to evaluate your improvements.

Result

Source code (including documentation) based on the current implementation of c3po. Description of the conflict resolution strategies based on tool priorities and other rules (e.g. if JHove claims the format to be X, DROID claims the format to be Y, then look for property Z: if present, JHOVE is right). Short quantitative analysis of the improvements achieved on the FITS data set (number of resolved conflicts, etc.). Structural description of the rule set.
For evaluation we will apply your version of c3po to a different set and evaluate the improvements independently.

Contact

For questions, contact Hannes Kulovits (kulovits@ifs.tuwien.ac.at)

3: Progressive content profiling (2x2)

Context

Content profiling has to achieve two contradictory goals: Efficiency of tools for identification and characterisation (http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/2452), but also in-depth feature extraction (http://ifs.tuwien.ac.at/~petrov/publications/c3po-poster-ipres12.pdf). These are not always compatible: there is a conflict between large-scale and in-depth profiling. Running FITS on all files of a large collection is hard (http://www.openplanetsfoundation.org/blogs/2013-01-09-year-fits); but only running *file* is not an option (http://www.openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits).

Goal

Your task is to combine more than one characterisation tool in c3po, to enable "progressive scanning" of collections. The idea is to enable a breadth-first characterisation that is combined with a drill down on specific parts. You will create a characterisation workflow that

- runs *file* on all items to retrieve the *mimetype*
- runs *Apache Tika* on selected items (e.g. all with certain mimetypes),
- runs *FITS* on all items of that set that have a certain property=value, and/or on mimetypes for which it works well and/or tika does not work well.

This requires you

- to develop a strategy and a mechanism for deciding which tool to run based on an output of a previous tool and a configuration file that allows customizing the workflow. You should use a Taverna workflow to control this.
- to develop a mechanism in c3po that allows you to combine property-value pairs from different sources. For example, you can develop a combined adaptor so that the different property-value pairs can be combined.

Discuss the speed improvements and the coverage of features that you can extract using this combination, and propose an optimal balance between detailed characterisation, medium level, and identification. What would be good strategies for balancing this?
Show how the values at different granularity and detail can be usefully combined to create a profile in c3po, and discuss the limits of this approach. What about conflicts?

Input

Please use the govdocs1 dataset which you can download from http://digitalcorpora.org/corpora/files. For this exercise we encourage you to use the 10 subsets of the entire corpus which you can download from http://digitalcorpora.org/corp/nps/files/govdocs1/zipfiles/ (thread0.zip ... thread9.zip). Each subset contains 1.000 files, randomly selected from the corpus, making up 10.000 files in total.
Fork from the c3po repository which can be found at: https://github.com/peshkira/c3po.

Concept

Discussion of the design of the c3po modules that are affected, and how they can be adapted. Comparison of the performance of *file*, *Apache Tika*, and *FITS*. Discussion of the key issues and how they will be tackled.

Results

Source code (including documentation) based on the current implementation of c3po. Prepare a live demonstration on the value of this using the govdocs1 corpus of data files, where all files are identified by *file*, roughly half is characterised by *Apache Tika*, and about 10% are characterised by *FITS*. Could the in-depth profiling be done on-demand? Which kind of infrastructure would be required for this?

Contact

For questions, contact Hannes Kulovits (kulovits@ifs.tuwien.ac.at)

4:Representative sets for experiments (2x2)

Context

Representative sample sets are required for experimentation in preservation planning. The sample files in their entirety should cover all properties present in the collection while trading off coverage against number of objects in the sample set. This, however, also might not always be sufficient since combinations of different properties in an object (encryption + proprietary file format, large number of pages + embedded images) may push migration and rendering tools to their limits. c3po provides an environment and interface for corresponding heuristics. However, the implemented heuristics in c3po don't meet a number of goals. Specifically, what is required for experiments is not necessarily a statistically fair sample of a data set according to selected dimension, but a test data set that covers the combinations of the feature set in a way that supports confidence in the tested behaviour of tools on the data.

Goal

Develop two additional heuristics for the representative set algorithm that achieve a configurable degree of coverage for a set of properties and values with a minimum number of objects (not necessarily a hard limit). Input to these heuristics is

- a set of property sets, associated where necessary with bin size configurations for numeric values. The sets can have size 1 (e.g. {{mimetype}, {validity}, {filesize}) or, in the advanced case, more than 1 (e.g. {{mimetype, validity}, {mimetype, filesize})
- A limit for the number of objects to be used, where 0 means no limit

Output is the set of objects, a coverage matrix (see below), and a measure of achieved coverage between 0 and 1, calculated as the percentage of covered cells in the coverage matrix.
Possible heuristics would include *simple greedy* (create a coverage matrix, i.e. a table with a row for each property and the possible values as cells, that maps out property-value combinations. Start filling up from the top and keep checking coverage until coverage is complete or the limit of objects is reached). One other heuristic is to be proposed in the concept.
Evaluate these heuristics for efficiency, speed and feature coverage on at least two different large data sets (one homogeneous, one heterogeneous), and discuss how realistic and representative the sets are. Start with sets of size=1 and then grow the sets a bit. How large are the resulting representative sets? Are you confident that you can get small sets that cover the "risk potential" of the data set? How can the heuristics be improved?
You can use one of your own data sets (this needs to be discussed with your tutor). Discuss the difference between statistical fairness and feature coverage and decide which is more important for DP experimentation and preservation planning.

Input

Please use the govdocs1 dataset which you can download from http://digitalcorpora.org/corpora/files. For this exercise we encourage you to use the 10 subsets of the entire corpus which you can download from http://digitalcorpora.org/corp/nps/files/govdocs1/zipfiles/ (thread0.zip ... thread9.zip). Each subset contains 1.000 files, randomly selected from the corpus, making up 10.000 files in total.
Fork from the c3po repository which can be found at: https://github.com/peshkira/c3po.

Concept

Describe in pseudo code the *simple greedy* and propose one other heuristic. Give some reasons to use the chosen heuristics including references to the literature you used. Discuss the c3po architecture and which parts of that architecture you will extend Explain how you will perform the evaluation.

Results

To submit:

- Source code (including documentation) based on the current implementation of c3po.
- Evaluation report

To present:

- Prepare a live demonstration, explain your own and the greedy heuristic and discuss how they perform.
- Explain whether the sample objects chosen by the algorithm increase your confidence in a preservation plan, decrease it, or not affect it?

Contact

For questions, contact Kresimir Duretec (duretec@ifs.tuwien.ac.at)

5: Interactive charting of multi-dimensional content profiles (2x2)

Context

Analysing large sets of files is a difficult and time consuming task. The content set addressed in a preservation plan is the result of a selection process based on the objects' properties. While characterisation tools such as *FITS* give an insight into the different properties of digital files (file size, image width/height, embedded metadata, number of pages, etc.), the content profiling tool *c3po* goes one step further by aggregating these properties. At present, c3po aggregates the metadata extracted from the digital objects independently from each other and presents this information as histograms to the user. This provides the user with a quick overview of the entire content set with its different characteristics. However, a combination of certain properties (e.g. file format and validity) is necessary to support partitioning of the collection into homogeneous sets based on these properties and thus enable an in-depth analysis of the objects characteristics.

Goal

Your task is to enhance c3po in a way that users can create bubble charts in a configurable way to visualise combinations of properties, e.g. file format (x) and validity (y). The third dimension necessary for the chart (size of the bubble) would in that case be the number of times a certain value (for validity either *true* or *false*) has been measured. The user shall be able to click a bubble in order to create the respective filter in c3po.

Input

Please use the govdocs1 dataset which you can download from http://digitalcorpora.org/corpora/files. For this exercise we encourage you to use the 10 subsets of the entire corpus which you can download from http://digitalcorpora.org/corp/nps/files/govdocs1/zipfiles/ (thread0.zip ... thread9.zip). Each subset contains 1.000 files, randomly selected from the corpus, making up 10.000 files in total.
For creating the bubble chart you can use jqPlot (http://www.jqplot.com). Please fork from the c3po repository which can be found at: https://github.com/peshkira/c3po.

Concept

Discussion of the design of the c3po modules that are affected, and how they can be adapted.

Result

Source code (including documentation) based on the current implementation of c3po. Prepare a live demonstration of your implementation using a set of files from the govdocs1 collection. Demonstrate in which cases (with which property combinations) this feature can be useful for analysing a collection.

Contact

For questions, contact Hannes Kulovits (kulovits@ifs.tuwien.ac.at)

6: Estimating the next migration

Context

Todays content owners host petabyes of data. Performing an action (migration) on such scale requires a lot of time for preparation, planning and in the end performing the action itself. To avoid bringing his content to a risk, a repository manager would like to have a possibility to estimate when will he need to migrate (or at least consider migrating) his content.
There are several reasons that can require considering a migration:

- content is getting too big (because of the ingest), so maybe there is an option to compress and save space and cost
- format support is fading (format is getting obsolete)
- users are not able to access the content for other reasons (such as features of content)

Those reasons can appear independently from each other at any point in time.

Goal

In this task you will build a simulation environment which will help to solve the above problem. You are free to use any technology/tool (MS Excel, Matlab, Octave, Java, Python, R ...) you like. Preferably, it should be open source but if you have a strong argument, you can use commercial product as well. (This needs to be agreed with your supervisor).
Your solution will provide a possibility to specify distributions for each reason: ingest, format obsolescence, access problems. Furthermore it will provide a possibility to specify a rule when a migration needs to be executed (considered). Based on those inputs your tool should output an estimation when is the most likely for a migration to appear.

Input

SIZE - current size of a collection
INGEST - distribution which defines the growth rate of a collection per month
FORMAT - distribution which defines the format obsolescence time
ACCESS - distribution which defines the time when collection won’t be accessible by the user community
RULES - rules when a migration should happen or at least it should be considered

Output

ESTIMATED_TIME - estimated time for the migration and probability
GRAPH - showing the distributions of migration times

Example

SIZE=100GB
INGEST=Normal(10GB,3)
FORMAT=Normal(10years,3)
ACCESS=Normal(5years,4)
RULES= migrate when size > 1TB or format is obsolete or access is not possible

ESTIMATED_TIME = October 2017 90%

Concept

In concept you should provide :

- decision which technology/tool to use and why (what are the benefits of using that technology)
- elaborate on type of simulation you will use (discrete event, stochastic , ...)
- which problems have you detected and how do you plan to solve them
- elaborate which distributions might be needed to cover format obsolescence and access

Result

As a result you should provide:

- implementation
- document describing your solution and results
- present a real world scenario
  - what are you estimations for the migration
  - how confident are you in those estimations
  - change the input parameters and observe what happens
  - what is the reason for a migration
- discuss which other aspects could be simulated

Contact: duretec@ifs.tuwien.ac.at

7: Business solution for preserving classical concert recordings

Context

The City of Vienna is determined to preserve every classical concert that is held within the city. Therefore it has contracted a company X to record every possible classical concert. The company X has been doing that for several years and already has 5TB of records in WAV format. Also their statistics show that the amount of data recorded per month is distributed with Normal distribution N(150 GB, 40 GB). The City of Vienna is now looking for a company which will **preserve** the recordings for the next 50 years. All companies are invited to submit their proposals by 06. 06. 2013. The City of Vienna will hire independent experts from the field of digital preservation and they will evaluate all proposals. The proposal should contain a detailed plan (not a Plato plan !) specifying all the activities that will be done to preserve the collection for 50 years.

The total cost are important factors but are definitely not the only one. The cheapest solution might not be chosen. Factors taught at digital preservation lectures at TU Wien should be also considered.

The City of Vienna will sign the contract with the best offer.

Goal

In this assignment you should write a proposal and offer your solution. Your proposal should cover all three levels of preservation : physical, logical and semantic (it is not necessary to provide preservation on this level). For each level (semantic can be excluded) you should provide analysis in terms of potential risks occurring, how you plan to cover them, and the resources (storage, human and financial resources) needed for doing that. For example, you could decide for the physical level to cover bit rot and hardware failures. To do that you could use redundant copies and checksums, but that will require more resources. Try to find as much real data as possible and where it is not possible try to make real assumptions. For the logical level you should also consider the fact that you are getting wav format: What could be the problems with that format? Which strategies will you use to reduce the risks? Are you considering the fact that the recording company might change the format for a better technology in the future?

Remember that you have competition, so try to be as realistic as possible. Also there might be a case where you can challenge other solutions or be challenged by other groups so it is important that you will be able to defend your results.

Input

No specific input

Concept

In the concept you should provide:

- business model canvas with short explanation of each element
- a list of risks you plan to cover

Result

As the result you should provide a complete specification defining :

- description of the system architecture
- analysis which shows why are you doing things that you are doing (why do you have 5 replicas instead of 2 and so on)
- description of digital preservation risks you cover and the way you cover them
- resource estimations - how much economical, technical and human resources you will need and what would be the costs for the City of Vienna

Useful article: http://ijdc.net/index.php/ijdc/article/download/143/205

Contact: **duretec@ifs.tuwien.ac.at**

8: Optimizing MCDM efficiency in preservation planning (2x1)

Context

In planning, a large number of decision criteria are evaluated. These have different weighting and utility functions. Only after filling out all of them does the decision maker see the composite effect. This is a quite expensive procedure, and the question arises if it can be done more efficiently. Can some of the criteria be left out from evaluation? Which? What is the effect of not providing a measure for them?

Goal

Your task is to develop an automated heuristic that takes as input a completed preservation plan and computes how many evaluation values could have been omitted in the evaluation step, without changing the result and threatening the confidence. The basic process for this can be like this:

- Create a copy of the tree with empty values
- Create an ordered list of the leaves of the tree according to criticality (1st) and impact (2nd)
- Start filling in values in the order of the list. After each value, check if one of the conditions is met
    - the alternative is rejected
    - the alternative cannot “win” anymore
- For rejected alternatives, do the remaining criteria need to be evaluated?
- Stop when the ranks of all alternatives are “decided” and the remaining unfilled leaves cannot change the final ranking. Calculate the savings.

Discuss the potential savings in the evaluation procedure. Furthermore, discuss at what risk these savings come. Could the trust of decision makers in the plans be affected? Is the increased speed worth the lack of information and evidence? What other options do you see to improve the efficiency?

Input

You will receive a completed plan to verify your procedure. We will test it on other plans too.
Plan: The XML representation of Digital Preservation of Console Video Games (SNES)
The Plato code base from https://github.com/openplanets/plato.

Concept

Step-wise description of the heuristic in pseudo-code
Outline of how to evaluate efficiency savings
Outline of cost-value-risk relation of key aspects when this is carried out (not the final discussion, but the outline of that discussion)

Result

Artifacts:

- Source code of the implementation on a private repository (there are plenty service providers, like https://bitbucket.org/.   Please add us as users .-)
- Report describing your solution, evaluation approach, and the discussion of your findings.

Demonstration:

- Open the preservation plan, step to Evaluate, and show those criteria that don’t have to be evaluated. Include a visual.
- Discuss trade-off decisions to be taken and limitations of this approach.

Contact:

For questions, contact Michael Kraxner (kraxner@ifs.tuwien.ac.at)

9: Data visualisation and analysis: Decision criteria (2x2)

Context

Decisions in preservation planning depend on a wide array of criteria. Some of these criteria can be easily measured automatically reducing the effort of the evaluation. Others, however, are either complex to obtain automatically or have to be obtained manually resulting in high effort in the planning process. Automation efforts are key to improve decision making efficiency - but which measures are most important to retrieve automatically? In order to support this decision, we need to benefit from previous plans as well as other decision makers by identifying criteria that are often used, have a high impact on the preservation planning result but are currently difficult to measure. Unfortunately, finding these criteria from the given data is not obvious. For more background information check the paper Improving decision support for software component selection through systematic cross-referencing and analysis of multiple decision criteria. To get more insight into criteria and how they are used in preservation planning you can look at existing public preservation plans in Plato.

Goal

Based on preservation planning data, create a visualisation that allows decision makers to identify those criteria and sets of criteria that are the best investment for automation development. The planning data is available in CSV format and contains the following dimensions:

- Category
- Attribute
- Measure
- Coverage (Frequency of occurrence)
- Impact
- Criticality
- Complexity of manual measurement (How difficult is it to measure this manually?)
- Scale of manual measurement (per Alternative or per Sample)
- Costs of automation development (How expensive is it to automate the measurement?)
- Costs of verifying automation
- Existence of automated measures
- Type of content

To highlight the criteria (-sets) that are often used but difficult to retrieve, create a bubble chart using the JavaScript visualisation library http://www.jqplot.com/. For example, the chart could show the following data:

- x-axis: Coverage
- y-axis: Cost of manual measurement (complexity * scale)
- Bubble size: Impact (or aggregated impact for sets)
- Bubble colour: Cost of automation (development + verification)

Allow for interactivity with the data (e.g. drill down in a criteria set to show the individual criteria). Define a suitable data format for the visualisation method (e.g. the JSON structure needed by jqplot) and prepare the data. Embed the implemented chart into a simple web page.

Think about ways to group the criteria into criteria sets. Plato's knowledge browser already has some criteria sets defined. Analyse the data using your visualisation method and summarise your findings. What are the low hanging fruit?

Input

You can donwload the data from https://tuwel.tuwien.ac.at/mod/resource/view.php?id=141594 in CSV format (see Goal).

Concept

Specify how your visualisation of criteria sets and criteria will look like.Define the data structure needed for the visualisation and describe it shortly. Keep in mind, that interactivity must be possible. Create a draft of the page containing example data and include a first version of the diagram.

Result

Please provide a zip file containing the report (pdf), the source code of your visualisation and instructions how to set everything up (if needed). Make sure your code is readable and include documentation and comments.
In your report, shortly describe the visualisation and your implementation. Analyse the provided data and summarise your findings. What criteria would you automate and why?

In the presentation, show your visualisation (with short demo), describe what value the visualisation has for a decision maker and present your findings from analysing the data.

Contact: For questions, contact Markus Plangg (plangg@ifs.tuwien.ac.at)

10: Quality assurance (2x2)

Context

Often, the result of preservation planning is to migrate files to another format - especially for non-interactive content (e.g. audio, images, videos, documents). To decide which tools to use and check the results, we have to do quality assurance to compare the migrated data to the original. Such QA is a data processing task that can benefit much from a workflow environment.
Taverna is an open source Workflow Management System – a suite of tools used to design and execute scientific workflows. It allows for the automation of experimental methods through the use of different (local or remote) services from a very diverse set of domains – biology, chemistry and medicine to music, meteorology and social sciences.
Taverna is also increasingly used in digital preservation to make migration, characterisation and quality assurance workflows. To allow programmatic execution, chaining and discovery of these workflows they have to have necessary metadata and a well-defined interface. Your task is to create such a workflow that allows to do quality assurance on files.

Goal

Create a workflow that takes two file paths and necessary parameters as input. Add quality assurance tools to the workflow. Most tools will output (semi-) structured data. Parse this data to retrieve five relevant measures and output each measure on a separate port.
You can find already defined measures in the preservation related RDF ontology at http://purl.org/DP/quality/measures. The measures you select depend on the type of file you want to do QA on. The ontology can be extended, if you find a measures that make sense in the context of quality assurance (e.g. test some property that is problematic during migration) and is missing.
Experiment with your QA workflow and, for each measures, try to find at least one test case that fails. Report on your findings.
The Taverna Component plugin allows to import Component Profiles that define an interface and necessary metadata to execute and query the workflow. You can find the relevant QA profile at https://github.com/openplanets/scape-component-profiles/tree/master/profiles. Metadata is added to the workflow as free-text or semantic annotations. The semantic annotations are based on RDF and expressed using the vocabulary at http://purl.org/DP/components#. Once your workflow is in a usable state, create a component from it. The plugin will do (very basic) validity checks but make sure to check it manually.
At the end of the exercise, publish the component to myExperiment.org.

Input

- You can retrieve the RDF ontology containing the currently defined measures from http://purl.org/DP/quality/measures that specify the output of the quality assurance algorithm.(You can find the complete vocabulary on https://github.com/openplanets/policies/tree/master/DP .)
- Further you'll receive the component profile that specifies the interface and required metadata of the workflow https://github.com/openplanets/scape-component-profiles/tree/master/profiles.

Concept

The concept for your exercise should contain a description of your test data and the measures you want to provide as a result of your quality assurance workflow. Also include a first draft of your workflow (it does not have to be fully working but should contain the basic structure you want to use).

Results

Please provide a zip file containing the report (pdf), the workflow(s) you created (.t2flow), the test results (Taverna results as XLS or output directory) and if possible the data used for testing (and/or a content profile of it). In your Report, very briefly describe your annotated workflow. Describe the measures you selected and implemented and explain why they are suitable for QA for your files. Describe the test data you used in your experiments, and how they were migrated. Discuss the findings of the experiments. What were the problematic properties? What measures failed for what files? Is there an underlying pattern?

In the presentation, shortly show your workflow, what measures it extracts and why they are suitable for quality assurance. Show your experiments and the results from your testing.

Further information

Setup

- Download and install Taverna Workbench from http://www.taverna.org.uk/.
  Please note that under Linux and Mac OS you also have to install graphviz as stated in the system requirements. It is used by Taverna to draw the workflow diagrams.
- Install the Taverna Component plugin from the update site http://build.mygrid.org.uk/taverna/internal/scape/240/.
- Set up tools needed for the quality assurance. Make sure the tools are in the path and set environment variables if necessary.

Hints and references

- To call command line tools from Taverna, use the *Tool* service available in *Service templates*.
- Taverna provides tools to parse data (e.g. XPath)
- Taverna has the ability to reuse workflows as *Nested workflow* available in *Service templates*.
- Taverna does implicit list handling as described in the Glossary#list_handling. This means when feeding a list of values into a service that expects a single value, Taverna will call the service once for each list entry and construct a list with one entry for each service run as output.
- Taverna does not support native conditional branching yet. If you need this, use this workflow as a starting point.
- http://www.myexperiment.org/ is a platform for sharing workflows. You can find a variety of workflows there including the Taverna starter pack that contains some sample workflows for common tasks.
- You may need to increase the available memory in the Taverna startup script.

Contact: For questions, contact Markus Plangg (plangg@ifs.tuwien.ac.at)

11. Resilient Web Services for Preservation of Business Processes - Framework implementation (2 x 2)

Background

Preservation of complete (business) processes, i.e. the software, hardware, data, etc. involved in the process, is a current and upcoming area of Digital Preservation research.
Processes consist of a series of processing steps that are executed in a specific order. They are highly complex and dynamic forms of digital objects.
Processes in a Service Oriented Architecture are consisting of discrete, loosely-coupled services which interoperate using a network and have clearly defined communication interfaces. Web services (WSs) are the most common way to realise SOA services.

While WSs bring a wealth of new possibilities and flexibility to business and scientific processes, they also introduce new risks for the process execution, such as a WS hosted by a third party becoming unavailable, which can bring the execution of the process to a halt, or the WS changing its communication interface, which may cause short downtimes in process execution, until the changes will be adopted into the process.
Finally, the behaviour of the WS may change, while the interface stays the same. This threat is extremely hard to detect, as the process may not break, but will still deliver the outputs, which, however, might not be correct, or different from what is expected. Such a type of change may happen due to many reasons, e.g.

- side-effects changes in the implementation of the webservice itself (e.g. code refactoring that changes the execution)
- bug-fixes in the webservice implementation
- changes in the dependency of the webservice implementation, e.g. a new version of a third-party library that computes different results, or in the hardware if some computation is done hardware specific (e.g. GPU computation)

What are the Resilient Web Services?

The concept of Resilient Web Services (RWSs) [Miksa, Mayer and Rauber, 2013] was introduced to ensure sustainability of any kind of processes which depend on them. RWSs provides a set of recommendations and guidelines on WS design that enrich standard WSDL specification with additional information and provide notification on changes in Web Services, e.g. change of timing characteristics (higher delays) or changes in functionality (more precise results of calculations). Resilient Web Services aim to provide information that would ease their long-term sustainability and usage. What kind of information needs to be provided is still an open research question. Methods include e.g.

- identifyYourself(), returning the current version number and auxiliary information such as last change date, determinism/statefulness

- identifySWEnvironment() and identifyHWEnvironment, returning the (essential) components of the software environment, such as the operating system (and version), libraries, and hardware environment, such as CPU/GPU model, ...
- serviceChangesSince(Date), swEnvironmentChangesSince(Date) and hwEnvironmentChangesSince(Date), returning a log of changes since the date provided.

Your task is to provide a prototype implementation of the framework to aid the transition from current webservices to resilient variants. You shall provide a basic implementation of these methods, for an operating system of your choice. This implementation should allow a providers of webservices to be used in an easy fashion to enrich their WS by these resilient webservice functionality. A convenient method would be to just have to override a basic class providing these methods already, and just adding the methods specific to the WS.
Your framework should peridodically, or upon request, check if something in the setup of the environment the webservice runs in has changed. Tools for gathering information on the environment on a specific point will be provided, you need implement a storage of this information, and a means to compare the environment at two different points.

Input

- some hints about Resilient Web Services
- For Linux
  - tools to capture hardware on Linux systems
  - tools to capture installed software packages on Debian based systems (via APT)

Results

- A skeleton that provides the methods required by Resilient Web Services on the OS of your choice, and which can be easily integrated into custom implementations of webservices.
- A demonstration of the usage of this framework, showing the functionality on a number of existing web services for which you can obtain the source code.

Hints and references

[Miksa, Mayer and Rauber, 2013]
Tomasz Miksa, Rudolf Mayer, and Andreas Rauber. Ensuring sustainability of web services dependent processes. International Journal of Computational Science and Engineering (IJCSE), 2013

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)

12. Resilient Web Services for Preservation of Business Processes - Register with Notifications (2 x 2)

Background

Preservation of complete (business) processes, i.e. the software, hardware, data, etc. involved in the process, is a current and upcoming area of Digital Preservation research.
Processes consist of a series of processing steps that are executed in a specific order. They are highly complex and dynamic forms of digital objects.
Processes in a Service Oriented Architecture are consisting of discrete, loosely-coupled services which interoperate using a network and have clearly defined communication interfaces. Web services (WSs) are the most common way to realise SOA services.

While WSs bring a wealth of new possibilities and flexibility to business and scientific processes, they also introduce new risks for the process execution, such as a WS hosted by a third party becoming unavailable, which can bring the execution of the process to a halt, or the WS changing its communication interface, which may cause short downtimes in process execution, until the changes will be adopted into the process.
Finally, the behaviour of the WS may change, while the interface stays the same. This threat is extremely hard to detect, as the process may not break, but will still deliver the outputs, which, however, might not be correct, or different from what is expected. Such a type of change may happen due to many reasons, e.g.

- side-effects changes in the implementation of the webservice itself (e.g. code refactoring that changes the execution)
- bug-fixes in the webservice implementation
- changes in the dependency of the webservice implementation, e.g. a new version of a third-party library that computes different results, or in the hardware if some computation is done hardware specific (e.g. GPU computation)

What are the Resilient Web Services?

The concept of Resilient Web Services (RWSs) [Miksa, Mayer and Rauber, 2013] was introduced to ensure sustainability of any kind of processes which depend on them. RWSs provides a set of recommendations and guidelines on WS design that enrich standard WSDL specification with additional information and provide notification on changes in Web Services, e.g. change of timing characteristics (higher delays) or changes in functionality (more precise results of calculations). Resilient Web Services aim to provide information that would ease their long-term sustainability and usage. What kind of information needs to be provided is still an open research question. Methods include e.g.

- identifyYourself(), returning the current version number and auxiliary information such as last change date, determinism/statefulness
- identifySWEnvironment() and identifyHWEnvironment, returning the (essential) components of the software environment, such as the operating system (and version), libraries, and hardware environment, such as CPU/GPU model, ...

- serviceChangesSince(Date), swEnvironmentChangesSince(Date) and hwEnvironmentChangesSince(Date), returning a log of changes since the date provided.

It takes some time to introduce a standard which would change the way the WSDL web services are specified. Furthermore, there is a wide range of WSs already operating, whose design will not be changed, because it involves effort from the sides operating them and this costs. For this reason, the easiest way to enhance the preservability of WSs is to decorate (design pattern) them, by providing additional information, which increases the preservability of processes dependent on them. This decoration will be made by you.

What is to be done?

In the course of this task you will have to research and propose what kind of information should enrich the description of web services in order to make them sustainable. For example, the WSDL description of the web service could be extended by the 'expiration date', which could be resolved in order to check how long the WS will be available. Having identified these information, you will have to implement an application, which would transform a web service into resilient web service, i.e. the application will receive the address of the WSDL description of the original web service, it will generate a new WSDL which will consist of the original methods and the 'resilient methods'. The 'resilient methods' will point to the web service implemented by you, which will provide 'resilient' information. Some of the 'resilient' information will be provided manually by the person registering the WS, while the other (and more important) will come from the monitoring of the registered WS, which your application will perform (tools for monitoring will be provided). When any kind of change to the monitored WS occurs, a notification will be sent through the 'resilient' methods.

Input

- some hints about Resilient Web Services
- WS monitoring tools

Results

- specification of Resilient Web Services
- web service with resilient information (provides resilient information about registered WSs)
- simple web register (paste link to the original WSDL, register and start monitoring, host the new WSDL)

Hints and references

[Miksa, Mayer and Rauber, 2013]
Tomasz Miksa, Rudolf Mayer, and Andreas Rauber. Ensuring sustainability of web services dependent processes. International Journal of Computational Science and Engineering (IJCSE), 2013

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)

13 Phaidra - assessment of the preservability of an archival system (2 x 2)

What is Phaidra?

Phaidra is a digital asset management system with long-term archiving functions, offers the possibility to archive valuable data university-wide with permanent security and systematic input, offering multilingual access using metadata, thus providing worldwide availability around the clock. As a constant data pool for administration, research and teaching, resources can be used flexibly, where continual citability allows the exact location and retrieval of prepared digital objects. The system is currently in use at Uni Wien. More: https://phaidra.univie.ac.at/

What is preservability?

Preservability is defined as the degree to which a system, product, or component can be archived for as long as necessary, ensuring its trustworthiness, and redeployed and re-executed according to the expectations, in a future environment, that might potentially be different from the original. This definition hints at the fact that the degree of preservability is always dependent on the requirements of the stakeholders, or in other words, it is dependent on the specific scenarios approached. Based on this definition, one can say that preservability seems to be a desired quality of systems, since it is usually not imposed by functional or business requirements.

What is to be done?

The aim of this task is to assess the preservability of the Phaidra system. The assessment will be conducted according to the assessment method developed in TIMBUS project. The method is compliant with ISO15504 for process assessment. An example of method application will be available. During the assessment, the group will have to identify important qualities of Phaidra system, which impact the preservability most. For this purpose ISO 25010 will be used as a basis. Goal modeling techniques and checklist assessment methods will be used to define and measure Phaidra's preservability capability.

Input

- virtual machine with an instance of Phaidra
- documentation of Phaidra
- assessment method
- example application of the method

Results

- assessment report presenting the evaluation of Phaidra in view of preservability
    - goal models

- checklist

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)

14 Validation and Verification of a migrated process (2 x 2)

Background

Preservation of complete (business) processes, i.e. the software, hardware, data, etc. involved in the process, is a current and upcoming area of Digital Preservation research.
Processes consist of a series of processing steps that are executed in a specific order. They are highly complex and dynamic forms of digital objects.

What?

An important step in digital preservation is on verification and validation that the preserved object still behaves the same as the original one. In this task, you shall thus deal with the verification and validation of a preserved business process.

We will provide you with a small and simple process that consists of a set of automated processing steps (w/o user input) for data analaysis, creating several intermediate outpus and finally creating a PDF report, containing tables and figures. This process is implemented using Latex, R, and other tools, and is originally deployed in a Windows environment.
You shall try to migrated this process (by migrating the software components and operating system environment used), and redeploy it in a new environment, either MacOS or Linux. There, you shall rerun the process. You shall come up with a plan to evaluate the correctness of the execution, and define measurements for this. Finally, perform this evaluation and determine if the process in the original and in the new environment behave the same.

The original process is run on a remote server, and the results can be obtained from a webservice. This webservice also provides you with the input data for your rerun of the redeployed process.

Input

- Input and output of the original process
- Description of a software setup that the process needs for execution.

Output

- The migrated process, in a packaged archive with instructions on how to run it.
- An evaluation of the execution of the migrated process

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)

15 Evaluation of processes - capturing measurements for significant properties comparison (2 x 2)

Background

Preservation of complete (business) processes, i.e. the software, hardware, data, etc. involved in the process, is a current and upcoming area of Digital Preservation research.
Processes consist of a series of processing steps that are executed in a specific order. They are highly complex and dynamic forms of digital objects.

What?

One important aspect in any preservation effort is the evaluation of the preservation actions, to verify that the significant properties are preserved in the modified object.
Dynamic objects require a slightly different approach for evaluation than traditional, more static objects.
An important part is the identification of external dependencies influencing the object's behavior. In the case of a process, this can be communication to external services, access to storage media, etc. Also the rendering of information on a screen, which would then e.g. be processed by a human, is an important aspect.
In this exercise, you shall thus focus on taking measurements relevant to evaluating whether two instances of a process are equal.

How?

Some ideas on what could be captured:

- Observing accesses to the file system; this can be done conveniently with the use of the jnotify java library, which provides you automatic event handling on file deletion, creation, ... See http://jnotify.sourceforge.net/ for more details.
- Observing the network traffic, using e.g. libraries such as jnetpcap, which uses low-level libraries as e.g. in wireshark. See http://jnetpcap.com/ for more details. You can also build upon a framework recently developed by us that employs jnetpcap to capture traffic specifically for webservices. To limit the amount of noise captured, you might also consider limiting listening to certain predefined destinations & ports.
- Capturing data on the resources consumed by the process, e.g. the duration, the (amount of) CPU and memory usage.
- Other ideas are of course welcome!

Note:For some of this capturing to be feasible, you might need to start the process in a controlled environment to obtain a PID.

You are free to reuse any tool that is freely available for this task, such as process tracing tools for Linux, Windows, etc..

As there will likely always be noise in the capturing (e.g. from other background processes running at the same time, e.g. virus checkers), you should foresee in your data model the possibility to run the same process several times, and store the capturing results for each run.

Then, some pattern recognition over this data should allow to identify which observations really belong to the process.

Input

- A sample process, implemented in two different fashions
    - As a Taverna workflow
    - As a shell script, calling various software modules
- Description of the process

Results

- Tools for capturing network traffic
- Tools for capturing file system operations
- An option to limit this capturing to ranges that are likely to be relevant for the process (e.g. listen to specific network ports only)
- A way to store the measurements, to allow for an easy comparison of different runs

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)

16 Automatic generation of a process context model instance from a workflow (2 x 2)

Background

Preservation of complete (business) processes, i.e. the software, hardware, data, etc. involved in the process, is a current and upcoming area of Digital Preservation research. Describing a process with all these aspects is an important first step towards preservation. To this end, a process context (meta)model that facilitates such a description of the various aspects and their relation to each other has been developed. The model covers aspects on business, application and technology (infrastructure) layers, e.g. information on persons, roles and services, applications, data, and hardware devices. For a specific process, an instance of this metamodel is created, describing the specific software and hardware setup, etc.
For processes that are already well defined in a workflow engine, the instantiation of this context model can draw on the workflow. This includes e.g. the software components, the data exchanged, etc.

What?

For this exercise, you shall work with the Taverna workflow engine; we will further provide you a specific workflow implemented, for which also a manually created context model instance is available. Your task is to use the Taverna API to read the structure of the workflow, extract all components identified, and potentially do further analysis on them. E.g. if you identify that a certain processing step uses a library, the license of this library can be identified using scripts such as licensecheck (http://manpages.ubuntu.com/manpages/lucid/man1/licensecheck.1.html). With this, you can capture the static description of the process.

In a second step, you shall also investigate in the runtime aspects of the process. Taverna already captures the data exchanged between steps during the workflow execution in a database, your task would be to extract this data and store it in the context model where fitting. Also, apply file-format identification tools to the data that is created during the workflow.

Input

- A sample process, implemented as a Taverna workflow
- A manually created instance of the context model, describing the process and its context and dependencies, as a reference point & example
- A description of the process

Results

- Tool for extracting information from a Taverna workflow into an instance of the context model
- An application of the tool on the sample processes provided

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Stefan Proell (sproell@sba-research.at)

17 Automatic recording and playback of user input for emulation evaluation in digital art and computer games (2 x 2)

Background

One important aspect of emulation is to evaluate that the original and preserved (emulated) environment still behave the same. For some types of systems, such as computer games, or digital art, user input is an important aspect. This input often controls the behaviour of the software. In many cases, the exact timing of when the input occurred is also important. Such user input can be in the form of mouse moves and clicks, keyboard input, etc.

To allow for an exact comparison of the original and emulated version of such a system, automatic recording and playback of this user input is vital.
Many emulation environments do not provide capabilities to record and/or later replay such user input. An alternative approach is thus the usage of virtual machines, which may provide an API to do such things.

In this task, you shall thus implement a prototype that shows the capabailities of the open-source virtual machine environment “Virtual Box”.
Virtual box provides an API that provides events for mouse and keyboard (thus enabling recording), and allows you to also generate events, thus enabling playback. Further, you can obtain the display of the virtual machine, and take screenshots, which enables the comparison.

Input

- Documentation of Virtual Box
- Examples of a digital art program and a computer game

Results

- A tool that can
  - record input in an Virutal Box image
  - playback that input (can be in the same image)
  - take screenshots at predefined events, likely before and after input events
- A report on evaluating the tool on a digital art and computer game software

Contact: for questions, contact Rudolf Mayer (mayer@ifs.tuwien.ac.at) and Tomasz Miksa (miksa@ifs.tuwien.ac.at)