Digital Preservation UE 2.0
188.475

Tasks 2011

Institut für Softwaretechnik und Interaktive Systeme
TU Wien

http://www.ifs.tuwien.ac.at/dp

Mode

- every task has a supervisor
  - who will have time for you
- submission of the concept on May 8th
  - explain the task in your own words and in more detail than the task description, so we can be sure you are clear about what you are supposed to do
  - what the expected results of the task are
  - how you plan to achieve these results
  - if there are choices (e.g. on a format/technology etc.) what you choose to do
  - what you already did as preparation for the task
  - your names / matrikelnumbers / Studienkennzahl of your study
  - concept will be part of the final grade for the course (20%)
- final submission on June 21st
  - written report
  - presentation (June 27th)
  - software, proper documentation
  - problems you encountered
  - findings from working on the task
  - literature with proper citations

Timeline

- 4.4. Tasks online
- 11.4. Presentation of tasks in the DP lecture, registering for tasks possible from after the lecture (in TUWEL)
- 15.4. Deadline for registration for tasks
- 8.5. Submission of concept papers
- 21.6. Submission of final reports
- 27.6. Presentation of final results

Persons: 1-3 Groups: 2 becker@ifs.tuwien.ac.at

- Some aspects of QA are and will remain HITs (Human Intelligence Tasks).
  - It is extremely difficult for a machine to decide if a certain property is "the same" on an intellectual level, when the logical representation is different
  - We can use crowd-sourcing to decide that
- Your Tasks:
  - Analyse how such a HIT would have to be specified to be understandable, doable, realistic and of the right granularity for a crowd-sourcing task
  - Investigate interfaces of crowd-sourcing platforms (Amazon Mechanical Turk, CrowdFlower…)
  - Specify the necessary "job submission"
  - Set up a crowd sourcing experiment for evaluating the quality of a certain migration experiment
    - Paid (costs will be paid by the university)
    - or un-paid (game design)
  - Assess results

Automated generation of benchmarking data with Office

Persons: 2-3 Groups: 3

becker@ifs.tuwien.ac.at

- Preservation characterisation and quality assurance, i.e. the analysis and comparison of properties of objects represented in certain formats, is very difficult without knowing what is in an object in the first place.
  - Q: When two measures contradict, how can we know which is true?
  - A: We can start from the other side and create objects ourselves.
- One challenge is: how can we generate the representation of these test objects in their formats? This should be done with the original environment, i.e. using Macros to simulate a human user creating a document.
- Your tasks are to create
  - A model (a simple domain-specific language, potentially in XML) describing properties of objects
  - A macro simulating a user creating (one/two of)
    - documents (Word)
    - spreadsheets (Excel)
    - presentations (PowerPoint)
    - shapes, diagrams, images (Visio)
    - images (Adobe Illustrator?)
    - CAD drawings?
    - something else? (Suggestions welcome)

becker@ifs.tuwien.ac.at

- Preservation characterisation and quality assurance, i.e. the analysis and comparison of properties of objects represented in certain formats, is very difficult without knowing what is in an object in the first place.
  - Q: When two measurement devices deliver contradicting values, how can we know which functions correctly?
  - A: We can start from the other side and create test objects ourselves.
- Map intellectual properties ("there is a table with 3 columns and 4 rows and no gridlines") into a format representation ("html ", Latex "\tabular", Word "tabs" or "table" or "embedded Excel object").
  - Choose one class of object (email, document, website...)
  - XML model of basic properties for that class of object
  - XML model of basic properties for at least 2 formats corresponding to this
  - mapping between intellectual and format-specific properties (how can intellectual property X be "implemented" in formats/tools A,B,C?)
  - sample models for object instances
  - automated transformation demo from intellectual to format-specific models

Automated generation of benchmarking data (web archiving) with XML and XSLT

Persons: 1-3 Groups: 2 becker@ifs.tuwien.ac.at

- Preservation characterisation and quality assurance, i.e. the analysis and comparison of properties of objects represented in certain formats, is very difficult without knowing what is in an object in the first place.
  - Q: When two measures contradict, how can we know which is true?
  - A: We can start from the other side and create objects ourselves.
- Generate the representation of these test objects in their formats
  - For HTML, we can use XSLT directly to create XHTML from the XML model.
- Tasks
  - An XML description (a simple domain-specific language based on XML) of atomic elements of web pages
  - XSLT generation of HTML
  - Test data suite

Automated generation of benchmarking data (databases) with XML and XSLT

Persons: 1-2 Groups: 2

becker@ifs.tuwien.ac.at

- Preservation characterisation and quality assurance, i.e. the analysis and comparison of properties of objects represented in certain formats, is very difficult without knowing what is in an object in the first place.
  - Q: When two measures contradict, how can we know which is true?
  - A: We can start from the other side and create objects ourselves.
- One challenge is: how can we generate the representation of these test objects in their formats? For SQL databases, we can use XSLT directly to create the SQL code from the XML model.
- Your tasks are to create
  - An XML description (a simple domain-specific language based on XML) of atomic elements of databases
  - XSLT generation of SQL code (for different database environments?)
  - Test data suite

Email preservation

Persons:1-3 Groups: 2

becker@ifs.tuwien.ac.at

- Emails often represent the most valuable repository of documents a person (or organisation) has
- Your task here is to
  - create a preservation plan for an email archive (e.g. MBOX)
  - analyse the degree of automation that can be achieved
- Depending on group size and interest, the focus could be more on the analysis and evaluation or the automation side.
- Essentially, the following tasks will arise:
  - Document scenario: users, time horizon, sample dataset...
  - Define requirements in objective tree
  - Specify measurable criteria in detail
  - Collect potential preservation actions (migration, emulation)
  - Evaluate actions
  - Analyse potential for improving automation: measurements, service integration
  - Assess results and recommend action

becker@ifs.tuwien.ac.at

- Preserv2 file format registry is designed to be a semantically enhanced registry containing information to aid in the process of digital preservation
- Data contained within this registry is pooled together from many other sources including data from The National Archives (UK) Pronom registry as well as dbpedia (semantically enriched wikipedia)
- Preserv2 registry provides open access to all the data contained within as well as services including a SPARQL endpoint and RESTful HTTP services.
- Data is currently available in XML and RDF formats.
- You will receive the complete RDF statements. Your task is to
  - review these and
  - extend the RDF structures defined in the P2 registry to cover additional formats, integrate the definitions of more formats, and/or to cover additional data sources.
- Preserve2:
  http://p2-registry.ecs.soton.ac.uk/

XML transcription using semantic technologies

Persons: 2-3 Groups: 1 becker@ifs.tuwien.ac.at

- several standards support the representation of numeric data sets:
  - CERIF, XFDU, NeXus, etc.
- a lot of valuable data are stored in structured form such as XML - but not necessarily in proper schemas (properly annotated data types, scales, ordering, etc.)

- Build an ontology of numeric data sets and their units, entities, measures
- XSLT transformation of numeric data sets
- Semi-automated annotation of numeric data with labels, units, etc.
- Transcription of structures according to configurable rule sets

- Exact tasks and focus will depend on group size and interests.

Preservation workflows (1)

becker@ifs.tuwien.ac.at

- Preservation workflows combine
  - Format identification
  - Characterisation
  - Action (e.g. Migration)
  - Analysis/QA
  - Reporting
- Taverna Workflow Engine
  - eScience processing workflows (data pipelines)
  - Used in bio-informatics and other fields
  - Open source, strong community
  - Powerful integration of services and code
  - Visually create workflows, debug, deploy to server, monitor execution...

Preservation workflows (2)

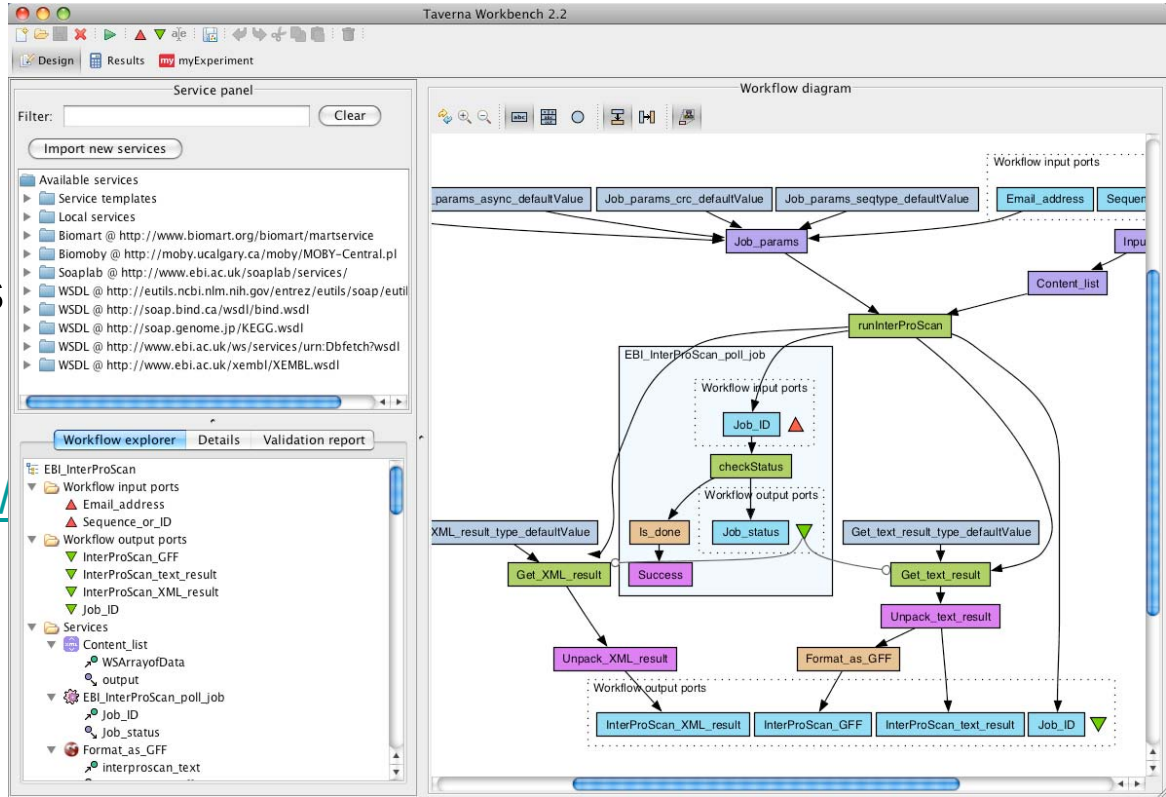
Persons: 1-2 Groups: 1-2

becker@ifs.tuwien.ac.at

- create 2-3 preservation workflows
  - format identification, basic property extraction, migration, report on results
- Run these on files
- Analyse
  - information needed about services
  - Connections (inputs and outputs)

http://www.taverna.org.uk/

http://www.myexperiment.org/



Migration web server Windows (Java)

Persons: 1-2 Groups: 1

becker@ifs.tuwien.ac.at

- Migration tools need to be discovered and invoked
- Discovery point and unified invocation methods in a service-oriented environment
- Task:
  - Create a simple registry describing migration services
  - Creat a simple service wrapper interface
  - Implement process measurement with a profiling tool like PsList
  - Configure and deploy 5 simple migration tools as services
  - Create a bootstrapping benchmark (depending on group size)

Migration web server Unix (Java)

Persons: 1-2 Groups: 1

becker@ifs.tuwien.ac.at

- Migration tools need to be discovered and invoked - discovery point and unified invocation methods to be useful in a service-oriented environment is required
  - Create a simple registry describing migration services
  - Create a simple service wrapper interface
  - Implement process management (forking, load monitoring)
  - Configure and deploy 5 simple migration tools as services
  - Integrate client QoS feedback (depending on group size)

Preservation of Office Formats

Persons: 1 Groups: 2 kulovits@ifs.tuwien.ac.at

- Microsoft's Office most commonly used environment for creating business documents
- particular challenges to DP: formats proprietary, objects composite
  - e.g. Microsoft Word document can contain different information types with different lifecycles; container remains accessible, specific component becomes impossible to render
- create one or more preservation plans given the following:
  - Thoroughly define planning context
  - Choose a well-defined collection of objects of a specific type, e.g. documents (MS Word), spreadsheet (MS Excel), or mails (MS Outlook)
  - Give an overall description of your chosen collection and describe the objects in the collection in detail
  - Define requirements in objective tree
  - Consider that documents may be electronically signed
  - Find and evaluate different alternative file formats

Preservation Planning: File Systems

Persons: 2 Groups: 1 kulovits@ifs.tuwien.ac.at

- evaluate different file systems with respect to their suitability for digital preservation type long-term storage
- create a complete preservation plan, focusing on elaboration of a comprehensive requirements tree comprising objectives to be met
  - Redundancy
  - error resiliancy
  - deployability of different target media
  - file size limitations
  - Configurability
  - suitability for small/large objects
  - Documentation
  - driver support
  - ....
- create the preservation plan using the planning tool Plato
  http://www.ifs.tuwien.ac.at/dp/plato
- create the requirements tree using the mind-mapping tool Freemind
  http://freemind.sourceforge.net

Technology watch

Persons: 1 Groups: 2

kulovits@ifs.tuwien.ac.at

- Create an application that monitors certain (sub-)domains on the web
- Similar to a web crawler, but only identify the files' format and keep log
- It shall fulfill the following requirements:
  - Keep record on the different file formats that occur and the number of files of that file format
  - Detailed specification of the file format. (Identifying files just from their file extension is not sufficient)
  - Send an alert when the number of files of a certain file format drop below a specified threshold (e.g. 10%)
- While other solutions are possible, this application might also be realized as a plug-in for the Heritrix web crawler.
- The concept already has to include a Software design

Simulator for BASIC programs of a simple Home computer

Persons: 2-3 Groups: 1

guttenbrunner@ifs.tuwien.ac.at

- write a simulator in Java that
  - is able to execute simple BASIC programs by stepping through the source code line by line and executing the instructions
  - produce the same graphical output as the original system (3 people working on the project)

- very basic system with simple commands
- very limited graphic abilities
- documents available (system documentation in various forms, format documentation from earlier DP task)
- various Source code files available, also the original machine to be able to test the programs
- code reuse (especially for graphics output) possible from an earlier DP project (migration of data from original system) (JAVA)

Preservation through Source Code Migration

Persons: 2-3 Groups: 1

guttenbrunner@ifs.tuwien.ac.at

- write a migration tool that
  - is able to convert simple BASIC programs to a non-obsolete programming language
  - language can be of your choice (JAVA, C++, Visual Basic or any other BASIC dialect, etc.)
  - produce the same graphical output as the original system (3 people working on the project)

- very basic system with simple commands
- very limited graphic abilities
- documents available (system documentation in various forms, format documentation from earlier DP task)
- various Source code files available, also the original machine to be able to test the programs
- code reuse (especially for graphics output) possible from an earlier DP project (migration of data from original system) (JAVA)

guttenbrunner@ifs.tuwien.ac.at

- interactive real-time presentations called "Demos" exist since the early 80's on almost all Computer and Video game system platforms
- preserve as part of the digital culture
- Your task is to think about
  - what metadata needs to be collected about these programs (for describing them but also for e.g. executing them again in an emulated environment)
  - what strategies are valid for preserving them
  - what are the significant properties that need to be preserved (-> preservation planning, create an objective tree, select appropriate sample objects, test your strategies against them)
- metadata and significant properties shall be valid for demos on all platforms -> necessary to look at different platforms (at least 3 fundamentally different systems like 8bit system, 16bit system, modern PC)
- Wikipedia article about Demo http://en.wikipedia.org/wiki/Demo_%28computer_programming%29

Preserving Virtual Worlds: Second Life

Persons: 2 Groups: 1

guttenbrunner@ifs.tuwien.ac.at

- Previous DP task was to record scenes in Second Life by controlling the avatar through Linden script or external tools, turned out to be not sufficiently flexible enough
- Extend open source Second Life viewer
- Your tasks are to
  - explore the possibilities of adding automated control for the avatar (either directly inside the viewer or by changing the interface so that external tools can be used)
  - explore the possibilities of extracting information from the virtual world to identify e.g. hot spots with either a lot of people or lots of user activity/interaction
  - explore the options to implement features for saving images/videos of the virtual world
  - dependent on the outcome of the explored possibilities create a concept for implementing features that allow an automated use of the viewer to move the avatar to sections of interest and record scenes at these sections
  - implement these features

Second Life:

http://www.ifs.tuwien.ac.at/dp/second_life/

Open Source Viewer:

http://wiki.secondlife.com/wiki/Open_Source_Portal

Emulation of old mobile devices

Persons: 1 Groups: 1

strodl@ifs.tuwien.ac.at

- mobile phones have become an integral part of people's daily lives, offer support for
  - Calendar, e-mails, applications, games, music
- number of operating system platforms from different vendors have been published allowing third party vendors software developments
- Your task is to
  - evaluate current approaches for emulation of mobile phones with focus on older generations
  - perform a structured evaluation of emulators against predefined criteria
  - focus of this analysis should be on the operating system and the support of third party vendor software (e.g. games, apps)
  - Are you able to demonstrate the phones functionalities in 10, 20 or even 100 years?
  - What are the current challenges?
  - What are potential preservation strategies for mobile phones and mobile applications.

Code Documentation Assessment

Persons: 1-2 Groups: 2

strodl@ifs.tuwien.ac.at
mayer@ifs.tuwien.ac.at

- Find out how well Java software is documented

- Your task is to
  - think about quality measurements for code documentation
  - develop of a software prototype that assesses existing source code
  - goal is the formal verification of the source code of a software system
  - what criteria need to be fulfilled by a useful documentation (a documentation that allows to understand the procedure and functionality of the system)?
  - can quality measurements be defined allowing to check the source code as a black box text?

Dependency Analysis of Digital Objects

Persons: 2 Groups: 2 mayer@ifs.tuwien.ac.at

- Correctly establishing the context of digital objects is an important aspect for correctly preserving their original meaning, semantic, and authenticity.
- One specific task is to identify dependencies between different incarnations of the same document, which might be different versions of the same document or different formats.
- dependency information can be discovered from meta-data and content.
- develop a prototypical framework that can identify these dependencies:
  - Assembly a corpus of test documents, containing (one class per group)
    - office documents
      - different versions: early drafts, versions with added paragraphs, versions with changes in previously existing text, versions with comments, ...
      - different formats: e.g. starting from a MS Word .doc format to a .docx, an open office version and a PDF export
    - manipulated images
      - cropped versions, resized versions, versions in different formats
  - Utilise tools to extract meta data from the objects
  - Utilise tools to analyse the content to detect e.g. image cropping or added comments
  - Develop a set of rules and heuristics to determine which objects might be in dependency to others
  - Store dependencies as RDF triples; think of a number of meaningful relation types for e.g. converted files, manipulated documents, different draft versions of the same document, etc.

Questions? Now is the time!

- Tasks:
  - Quality Assurance in the crowd (CB)
  - Model-driven engineering: transform intellectual into format-specific properties (CB)
  - Automated generation of benchmarking data with Office (CB)
  - Automated generation of benchmarking data (web archiving) with XML and XSLT (CB)
  - Automated generation of benchmarking data (databases) with XML and XSLT (CB)
  - Preservation workflows in Taverna (CB)
  - Email preservation (CB)
  - XML transcription using semantic technologies (CB)
  - Migration web server Windows (Java) (CB)
  - Migration web server Unix (Java) (CB)
  - Using semantic technology for file format risk assessment (CB)
  - Preservation of Office Formats (HK)
  - Preservation Planning: File Systems (HK)
  - Technology watch (HK)
  - Simulator for BASIC programs of a simple Home computer (MG)
  - Preservation through Source Code Migration (MG)
  - Preserving Digital Art (MG)
  - Preserving Virtual Worlds: Second Life (MG)
  - Emulation of old mobile devices (SS)
  - Code Documentation Assesment (SS, RM)
  - Dependency Analysis of Digital Objects (RM)
- TUWEL Registration from now until April 15th!