Preserving
Complex Scientific Objects:
Process Capture and Data Identification

Andreas Rauber

J.Binder, T.Miksa, R.Mayer, S.Pröll, S.Strodl, M.Unterberger

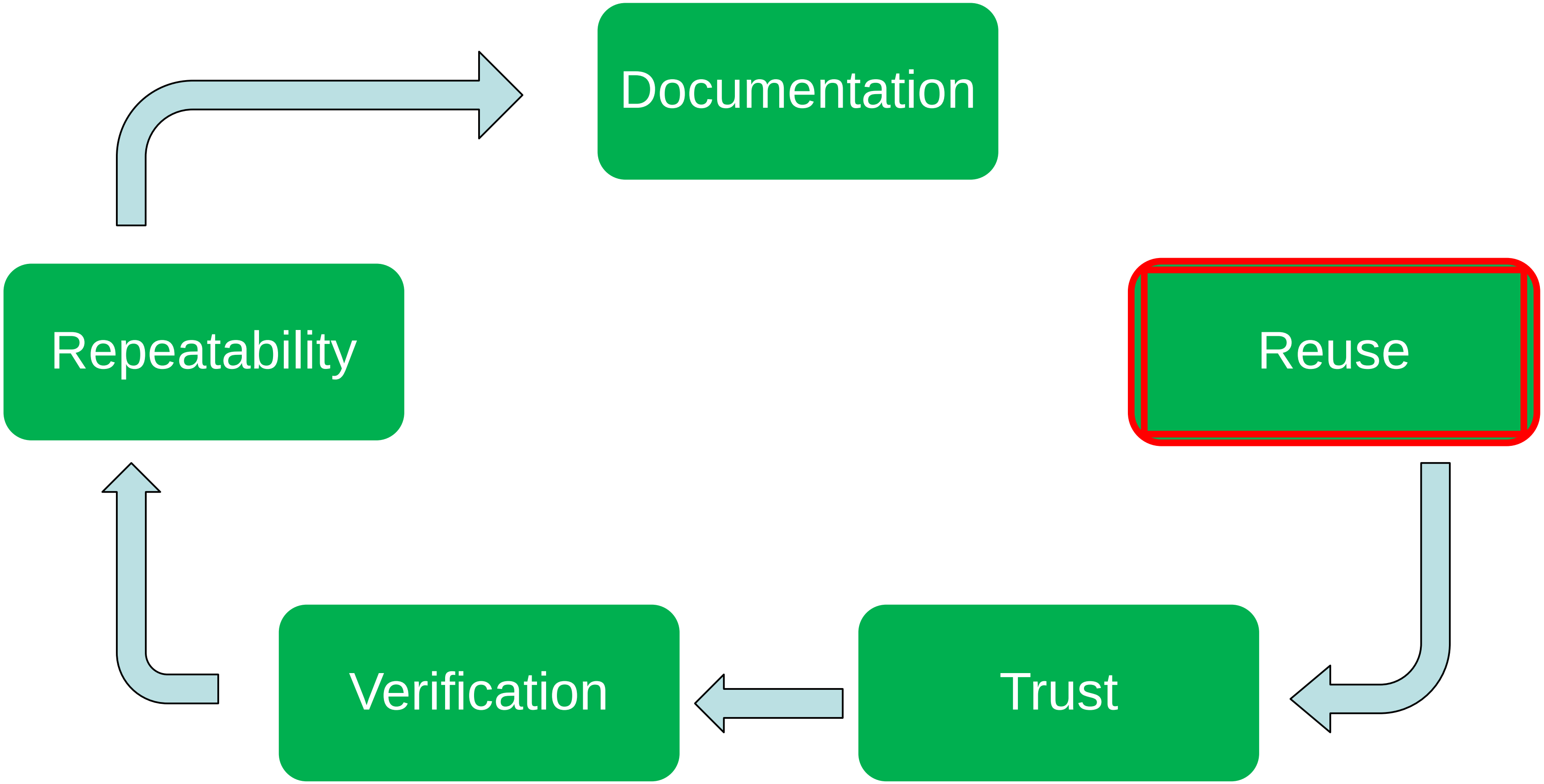
Vienna University of Technology
&
Secure Business Austria
rauber@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/~andi

Outline

- What is the “Complex Scientific Object” to preserve?
- How to capture a process and its context?
- How can we precisely identify the data used?
- Summary

Preserving Research

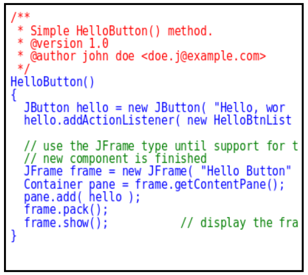
- Why do we want to preserve research/scientific objects?



Preserving Research

- Preservation:
  - "keeping useable over time"
  - fighting technical & semantic obsolescence

EVALUATION OF FEATURE EXTRACTORS AND PSYCHO-ACOUSTIC TRANSFORMATIONS FOR MUSIC GENRE CLASSIFICATION



```
/**
 * Simple HelloButton() method.
 * @version 1.0
 * @author john doe <doe.j@example.com>
 */
HelloButton()
{
  JButton hello = new JButton( "Hello, wor
  hello.addActionListener( new HelloBtnList

  // use the JFrame type until support for t
  // new component is finished
  JFrame frame = new JFrame( "Hello Button"
  Container pane = frame.getContentPane();
  pane.add( hello );
  frame.pack();
  frame.show();            // display the fra
}
```

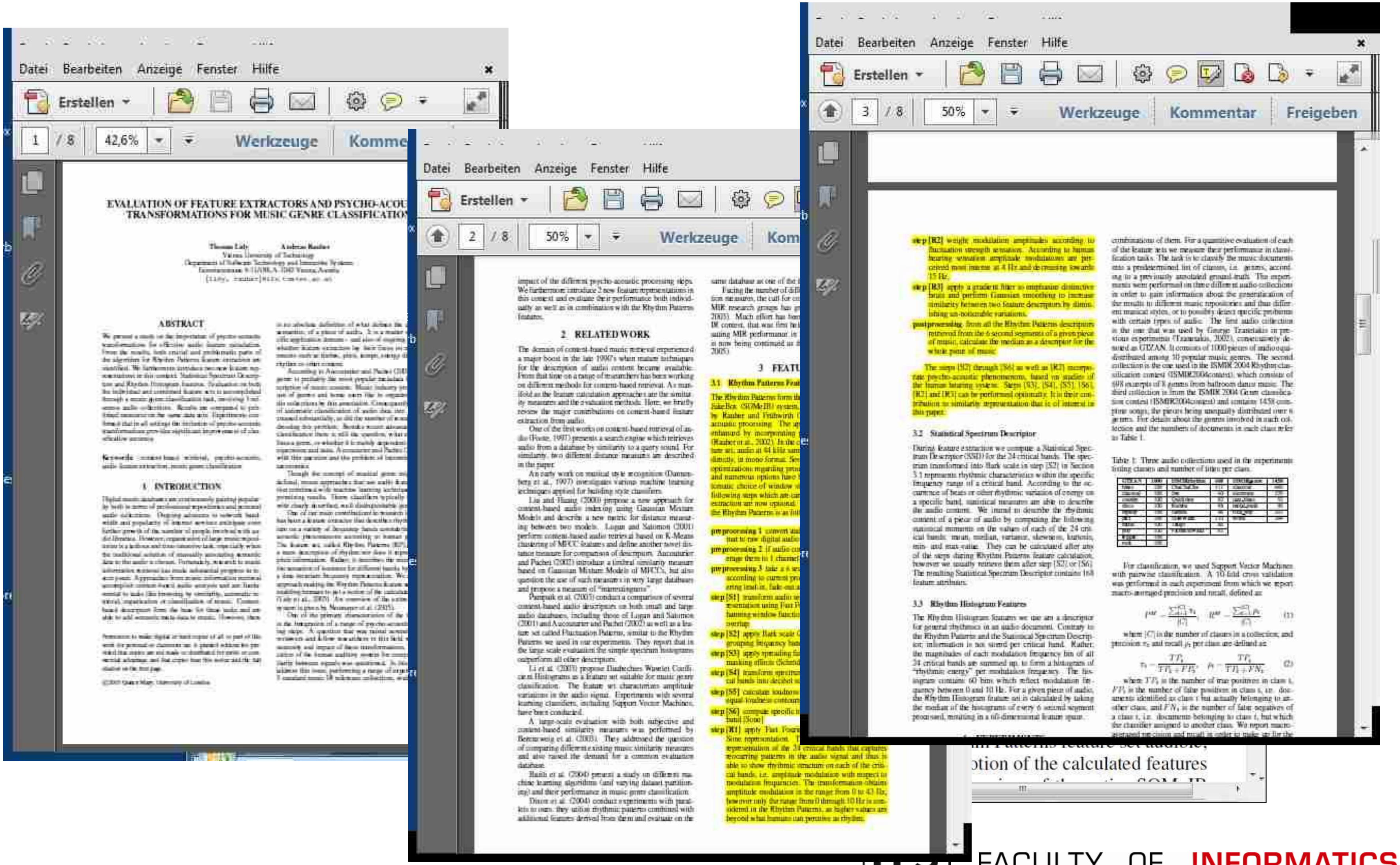
- Research:
  Which "Scientific Objects"

- What have we got?
  - Research Objects
  - Repositories for papers, data and code
  - Data Management Plans

Done!?

From Data to Processes

- Excursion: Scientific Processes



EVALUATION OF FEATURE EXTRACTORS AND PSYCHO-ACOU
TRANSFORMATIONS FOR MUSIC GENRE CLASSIFICATION

ABSTRACT

1 INTRODUCTION

2 RELATED WORK

3 FEATU

3.1 Rhythm Patterns Feat

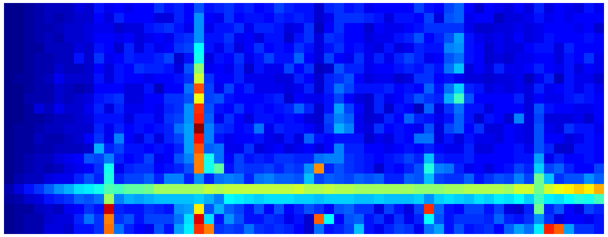
3.2 Statistical Spectrum Descriptor

3.3 Rhythm Histogram Features

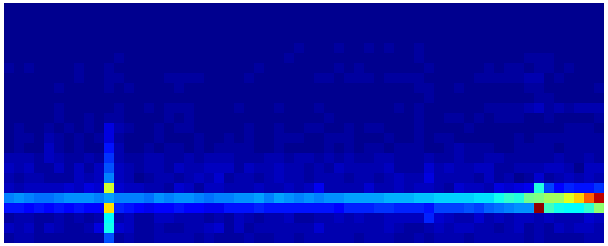
Table 1: Three audio collections used in the experiments listing classes and number of titles per class.

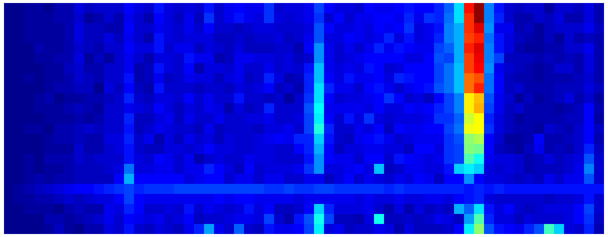
From Data to Processes

- Excursion: scientific processes

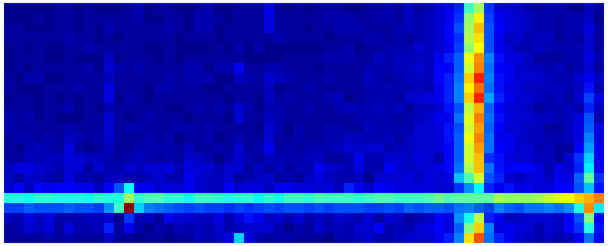


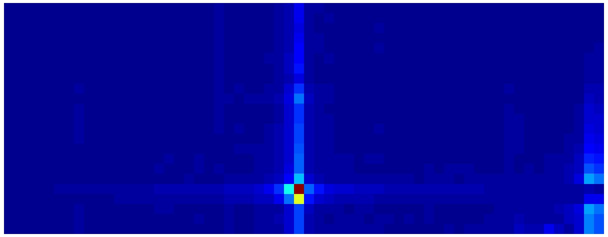
set1_freq440Hz_Am11.0Hz



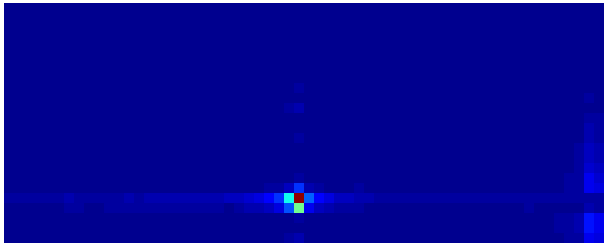


set1_freq440Hz_Am12.0Hz





set1_freq440Hz_Am05.5Hz

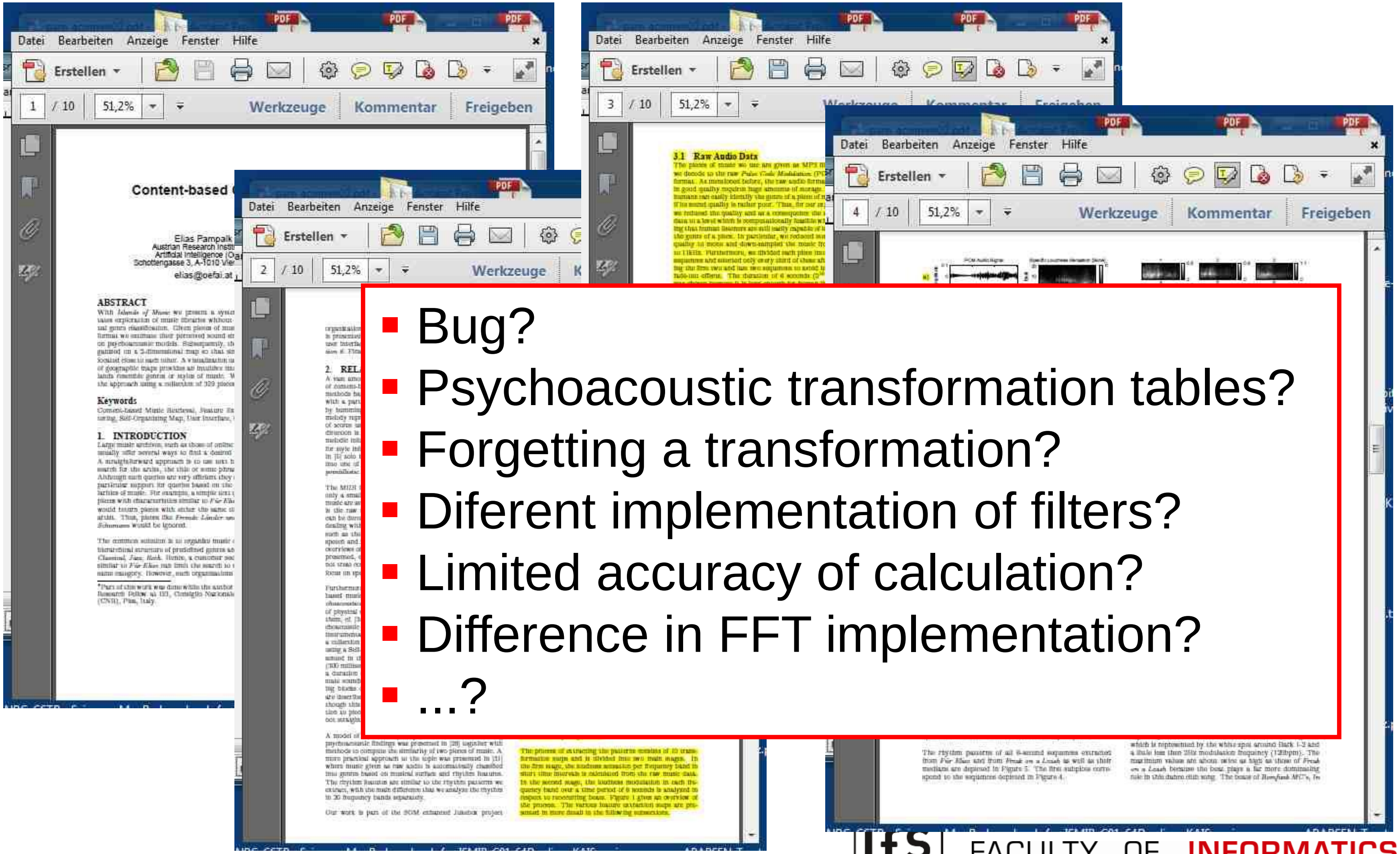


Java

Matlab

From Data to Processes

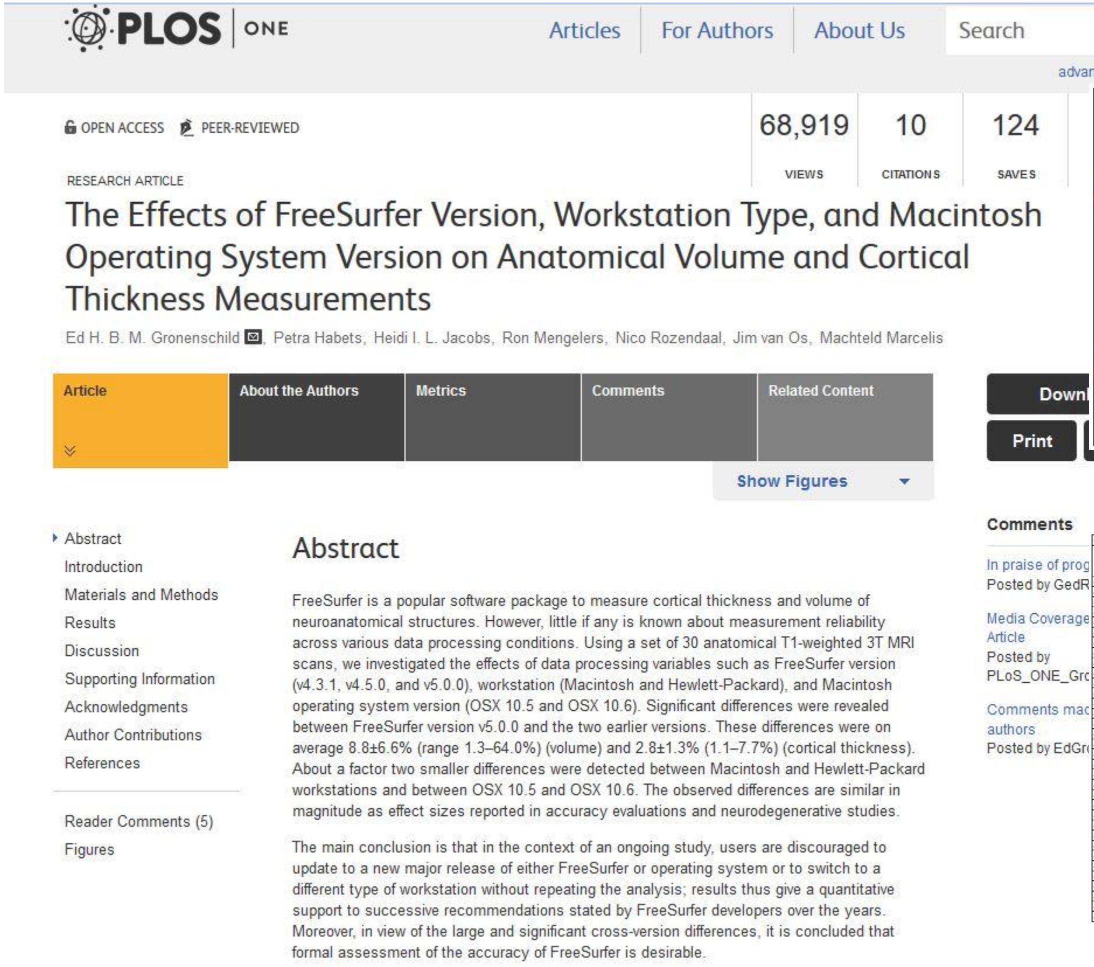
- Excursion: Scientific Processes



- Bug?
- Psychoacoustic transformation tables?
- Forgetting a transformation?
- Diferent implementation of filters?
- Limited accuracy of calculation?
- Difference in FFT implementation?
- ...?

From Data to Processes

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234



PLOS ONE | Articles | For Authors | About Us | Search | advanced search

OPEN ACCESS | PEER-REVIEWED

68,919 VIEWS | 10 CITATIONS | 124 SAVES

RESEARCH ARTICLE

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Article | About the Authors | Metrics | Comments | Related Content

Show Figures

Down | Print

- Abstract
- Introduction
- Materials and Methods
- Results
- Discussion
- Supporting Information
- Acknowledgments
- Author Contributions
- References

- Reader Comments (5)
- Figures

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

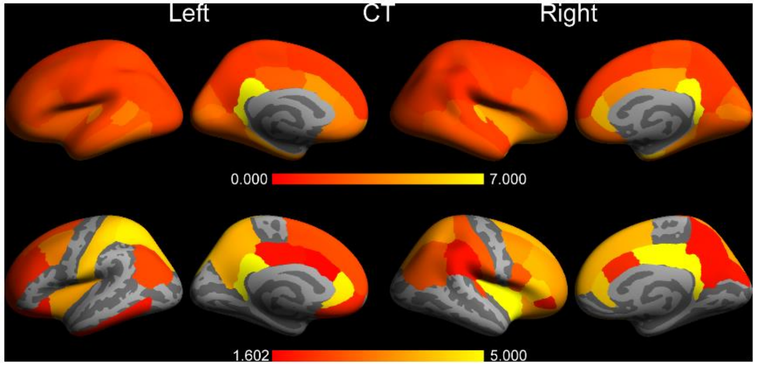
The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

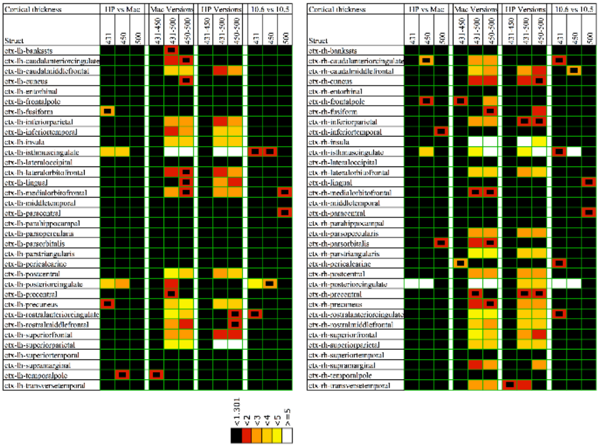
Comments

In praise of prog
Posted by GedR

Media Coverage
Article
Posted by
PLoS_ONE_Grc

Comments mac
authors
Posted by EdGr



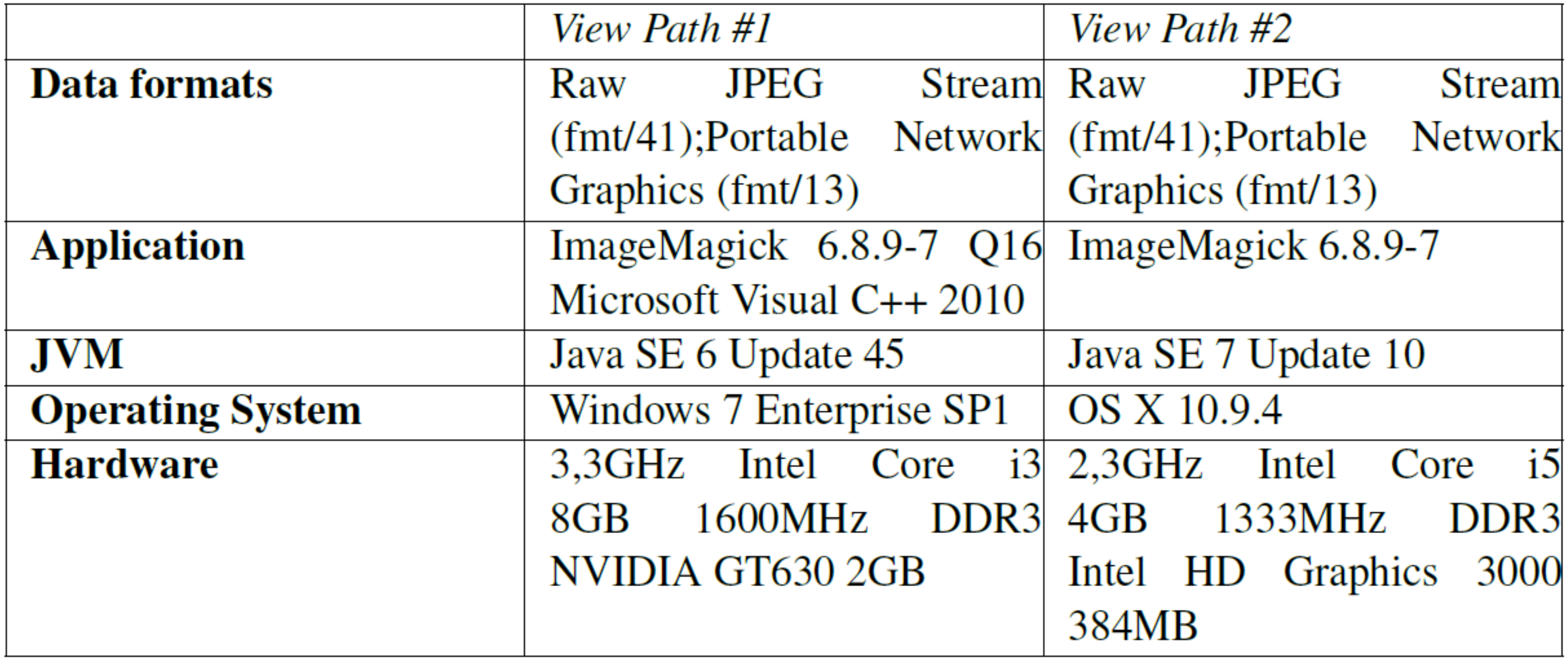


ifs FACULTY OF !NFORMATICS

From Data to Processes

A simpler example

- Image conversion from jpg to tiff using *ImageMagick*



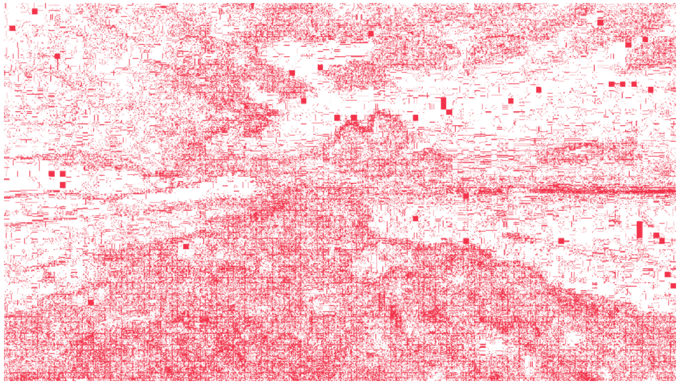
	View Path #1	*View Path #2*
Data formats	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)
Application	ImageMagick 6.8.9-7 Q16 Microsoft Visual C++ 2010	ImageMagick 6.8.9-7
JVM	Java SE 6 Update 45	Java SE 7 Update 10
Operating System	Windows 7 Enterprise SP1	OS X 10.9.4
Hardware	3,3GHz Intel Core i3 8GB 1600MHz DDR3 NVIDIA GT630 2GB	2,3GHz Intel Core i5 4GB 1333MHz DDR3 Intel HD Graphics 3000 384MB

From Data to Processes

Original jpg

TIFF
Migration on Windows7

TIFF
Migration on OSX



Diff

Process Management Plans

- Need to preserve the process, not (only) the outputs!

- **“Process Management Plans” (PMPs)?**
  - Go beyond data (DMPs) to cover research process:
    - ideas, steps, tools, documentation, results, …
    - data is only one (important) element,
      usually a result of a research (pre-)process
  - Ensure re-executability, re-usability
  - Should be machine-actionable & verifiable
  - Basis for preservation and re-use of research
  - Similar to “research objects”, “executable papers”, …
  - Should be created semi-automatically

Outline

- What is the “Complex Scientific Object” to preserve?
- How to capture a process and its context?
- How can we precisely identify the data used?
- Summary

Process Management Plans

Need to create

- Models for representing such "process management plans" (PMPs)
- Should be machine-readable and machine-actionable
- Identify "**minimum** set" of information
- Devise means to automate (most of) the activity in creating and maintaining those PMPs
- Establish them to replace (enhance / subsume / …) Data Management Plans

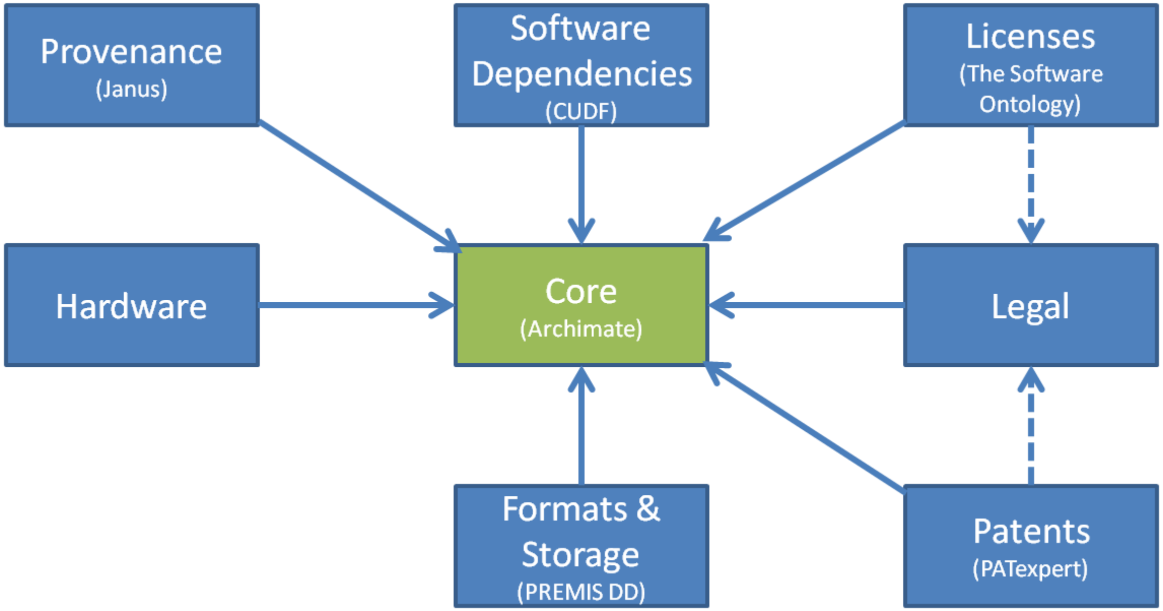
Process Management Plans

Structure of PMPs (following concept of DMPs):

1. Overview and context
2. Description of process and its implementation
   - Process description | Process implementation | Data used and produced by process
3. Preservation
   - Preservation history | Long term storage and funding
4. Sharing and reuse
   - Sharing | Reuse | **Verification** | Legal aspects
5. **Monitoring** and external dependencies
6. Adherence and Review

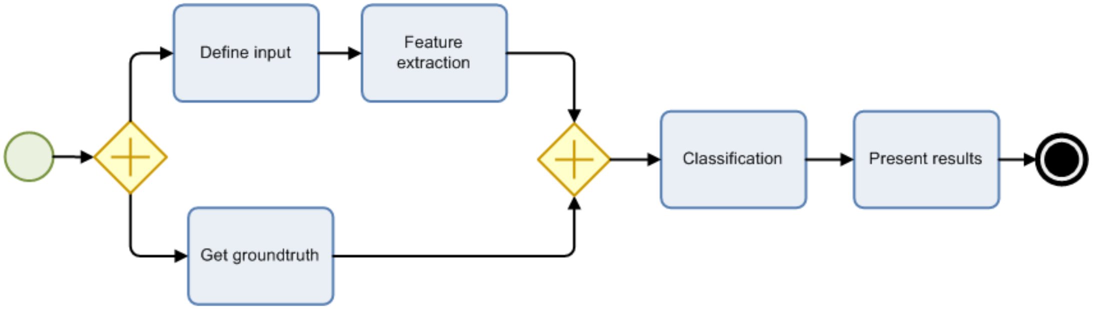
Process Context Model

- Establish what to document and how: Context Model
- Meta-model for describing process & context
  - Extensible architecture integrated by core model
  - Reusing existing models as much as possible
  - Implemented using OWL



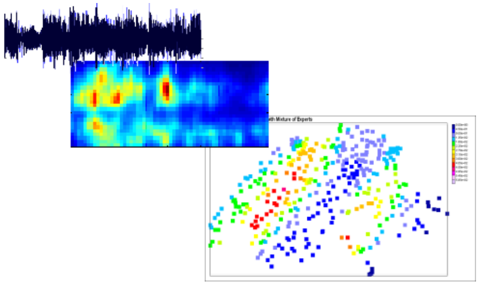
Provenance (Janus)
Software Dependencies (CUDF)
Licenses (The Software Ontology)
Hardware
Core (Archimate)
Legal
Formats & Storage (PREMIS DD)
Patents (PATexpert)

Application Example: Steps



- Acquisition of music & ground-truth data

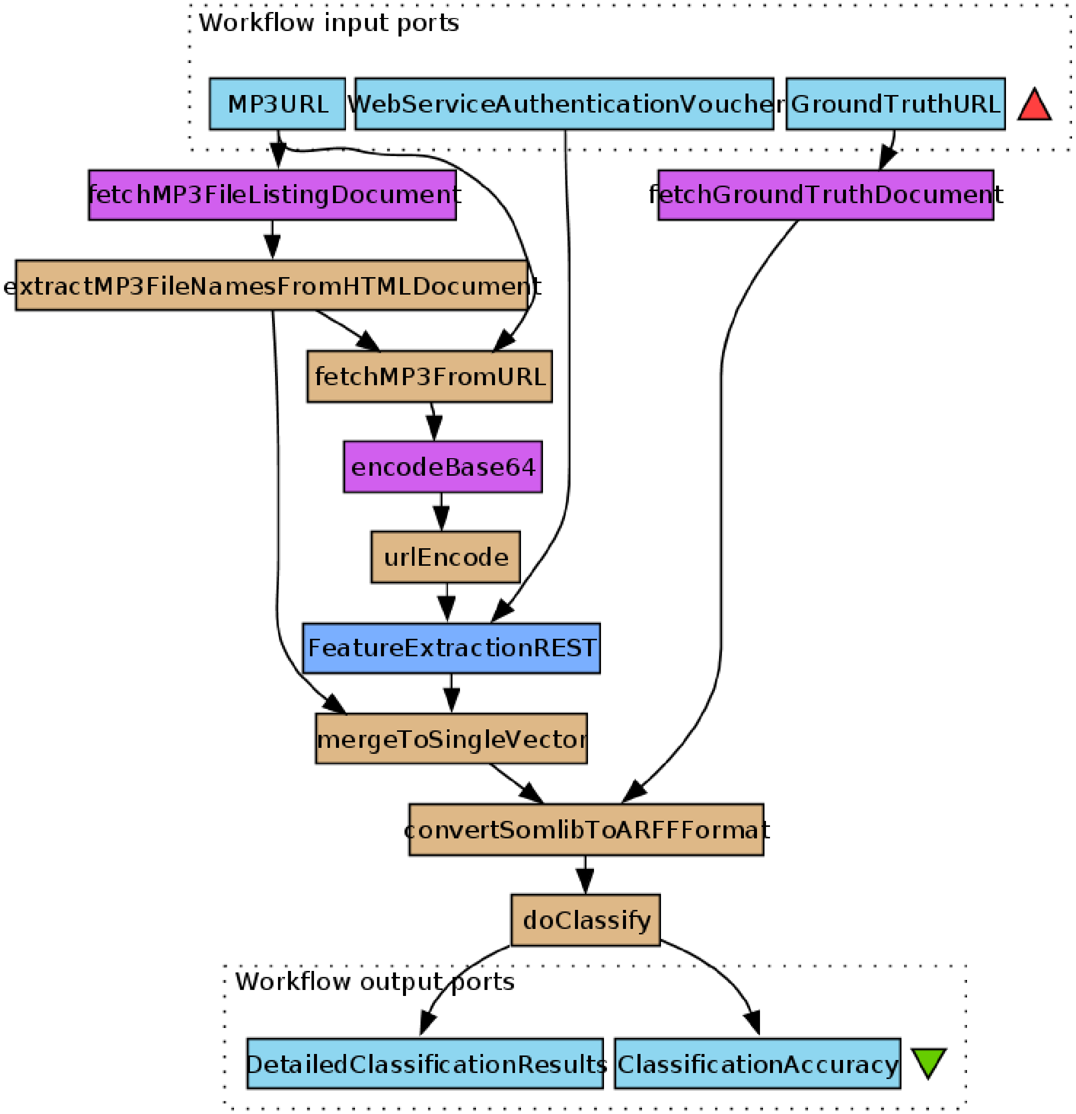
- Extraction of numeric features
- Training of machine learning model

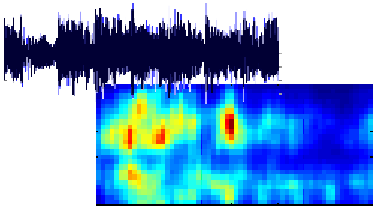


- Analysis of classification performance
- Repetition of experiment with variations
  - Finally leading to publication

Process Capture

Taverna



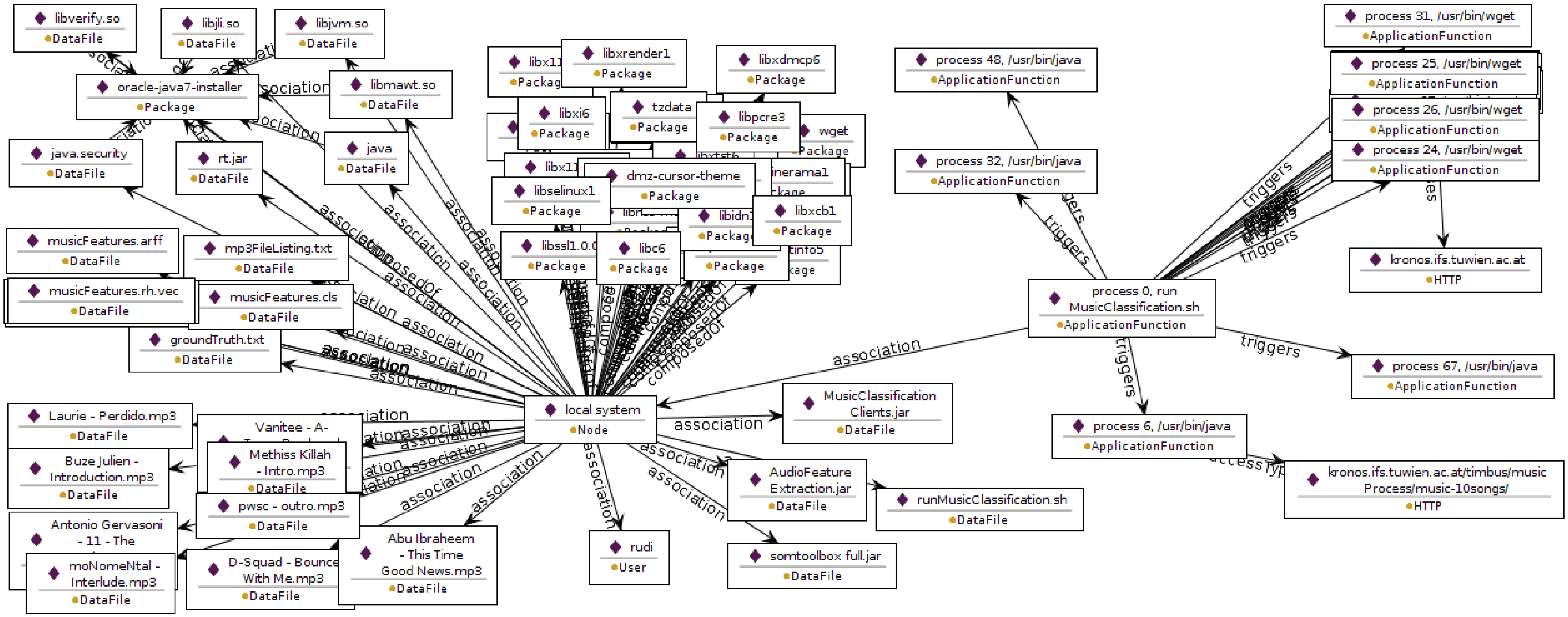


Kepler

Activiti

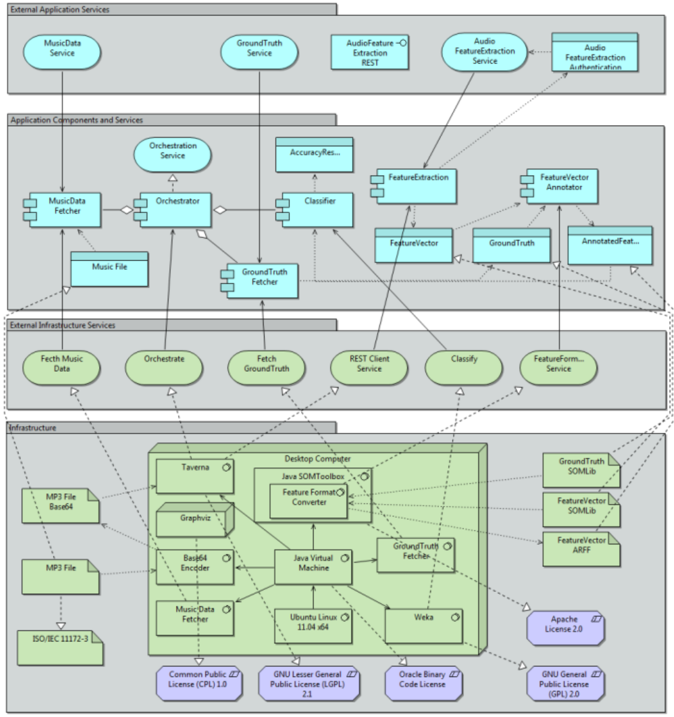
Automatic Model Generation

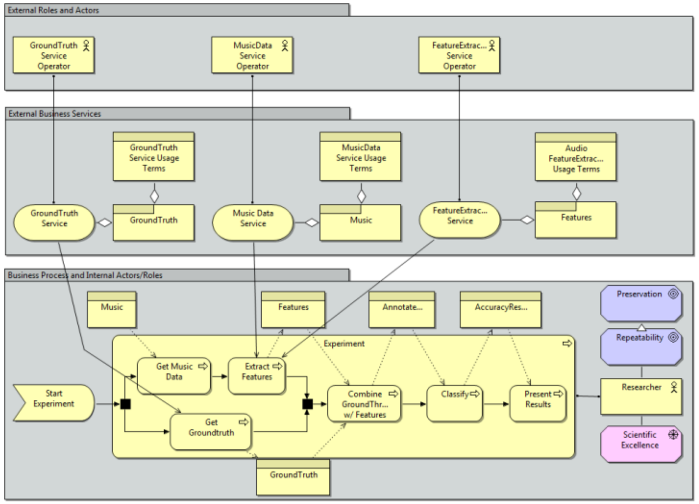
- Bottom up: tracing of specific execution
  - Captures all resources accessed (files, ports, ...)
  - Linux prototype (http://ifs.tuwien.ac.at/dp/process/projects/pmf.html)
  - Captures verification data of process execution instance



Automatic Model Generation

- Top-down: capturing of execution environment
  http://opensourceprojects.eu/p/timbus/
  - Software applications & dependencies
    (Linux Packages & Windows DLLs)
  - Licenses (mostly Open Source)
  - File Formats (DROID) &
    Link to registries (PRONOM)
  - Hardware (Linux & Windows)

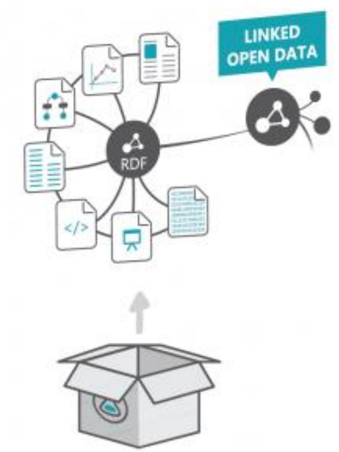


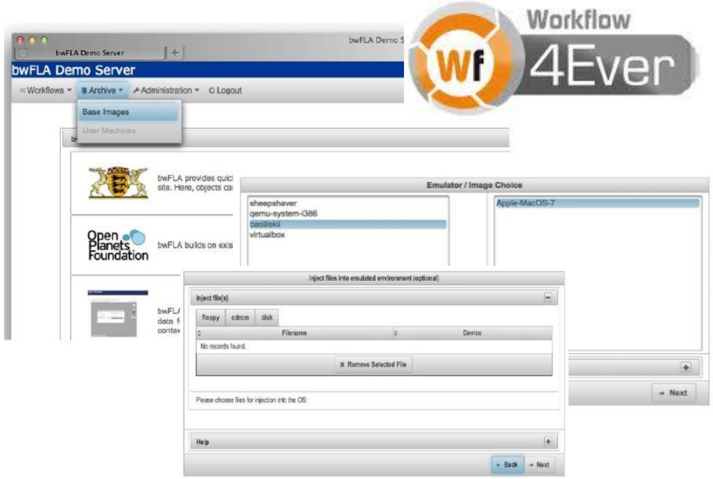

External Roles and Actors
GroundTruth Service Operator
MusicData Service Operator
FeatureExtrac... Service Operator
External Business Services
GroundTruth Service Usage Terms
MusicData Service Usage Terms
Audio FeatureExtrac... Usage Terms
GroundTruth Service
GroundTruth
Music Data Service
Music
FeatureExtrac... Service
Features
Business Process and Internal Actors/Roles
Music
Features
Annotate...
AccuracyRes...
Preservation
Experiment
Repeatability
Get Music Data
Extract Features
Start Experiment
Combine GroundThr... w/ Features
Classify
Present Results
Researcher
Get Groundtruth
Scientific Excellence
GroundTruth

Process Capture

Preservation and Re-deployment

- „Encapsulate“ as complex Research Object (RO)
- DP: Re-Deployment beyond original environment
  - Format migration of elements of ROs
  - Cross-compilation of code
  - Emulation-as-a-Service
- Verification upon re-deployment

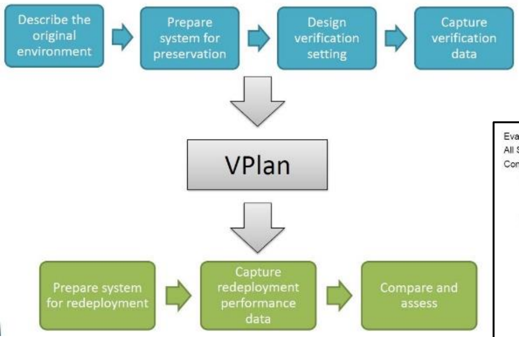


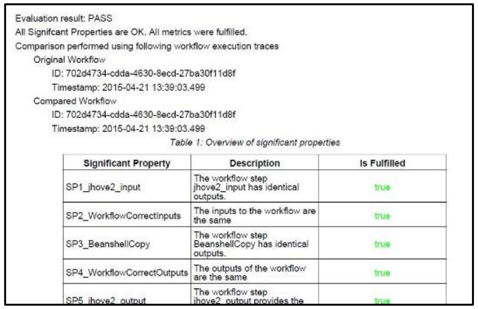


Emulator / Image Choice: sheepshaver, qemu-system-i386, basiliskii (selected), virtualbox — Apple-MacOS-7 (selected)

No records found.

Please choose files for injection into the OS


Describe the original environment
Prepare system for preservation
Design verification setting
Capture verification data
VPlan
Prepare system for redeployment
Capture redeployment performance data
Compare and assess



```
Evaluation result: PASS
All Signifcant Properties are OK. All metrics were fulfilled.
Comparison performed using following workflow execution traces
    Original Workflow
        ID: 702d4734-cdda-4630-8ecd-27ba30f11d8f
        Timestamp: 2015-04-21 13:39:03.499
    Compared Workflow
        ID: 702d4734-cdda-4630-8ecd-27ba30f11d8f
        Timestamp: 2015-04-21 13:39:03.499
```

Table 1: Overview of significant properties

Significant Property	Description	Is Fulfilled
SP1_jhove2_input	The workflow step jhove2_input has identical outputs.	true
SP2_WorkflowCorrectInputs	The inputs to the workflow are the same	true
SP3_BeanshellCopy	The workflow step BeanshellCopy has identical outputs.	true
SP4_WorkflowCorrectOutputs	The outputs of the workflow are the same	true
SP5_jhove2_output	The workflow step jhove2_output provides the	true

Outline

- What is the “Complex Scientific Object” to preserve?
- How to capture a process and its context?
- How can we precisely identify the data used?
- Summary

Data and Data Citation

- So far focus on the process
- Processes work with data
- Data as a “1st-class citizen” in science
- We need to be able to
  - preserve data and keep it accessible
  - cite data to give credit and show which data was used
  - **identify precisely the data used** in a study/process for repeatability, verifyability,…
- Why is this difficult?
  (after all, it’s being done…)

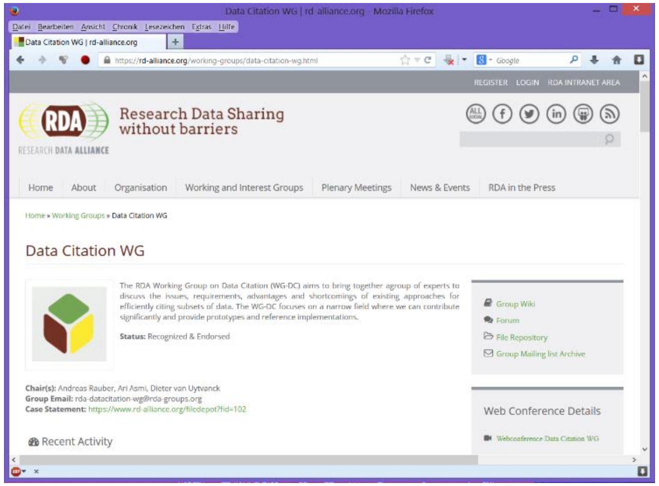
Granularity of Data Identification

- What about the **granularity** of data to be identified?
  - Databases collect enormous amounts of data over time
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire dataset, providing textual description of subset -> imprecise (ambiguity)
  - Storing list of record identifiers in subset -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to cite precisely the **subset of (dynamic) data used** in a study

Identification of Dynamic Data

- Citable datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
- But: (research) data is **dynamic**
  - Adding new data, correcting errors, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Current approaches
  - Identifying entire data stream, without any versioning
  - Using “accessed at” date
  - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to cite precisely the **data as it existed at any point in time**

RDA WG Data Citation


Data Citation WG | rd-alliance.org - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Data Citation WG | rd-alliance.org
https://rd-alliance.org/working-groups/data-citation-wg.html
Google
REGISTER LOGIN RDA INTRANET AREA
RESEARCH DATA ALLIANCE
Research Data Sharing without barriers
Home About Organisation Working and Interest Groups Plenary Meetings News & Events RDA in the Press
Home » Working Groups » Data Citation WG
Group Wiki
Forum
File Repository
Group Mailing list Archive
Web Conference Details
Webconference Data Citation WG
Recent Activity

Data Citation WG

The RDA Working Group on Data Citation (WG-DC) aims to bring together agroup of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Status: Recognized & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck
Group Email: rda-datacitation-wg@rda-groups.org
Case Statement: https://www.rd-alliance.org/filedepot?fid=102

- Research Data Alliance
- WG on **Data Citation:**
  Making Dynamic Data Citeable
- WG officially endorsed in March 2014
  - Concentrating on the problems of
    large, dynamic (changing) datasets
  - Focus! Identification of data!
    Not: PID systems, metadata, citation string, attribution, …
  - Liaise with other WGs and initiatives on data citation
    (CODATA, DataCite, Force11, …)

  - https://rd-alliance.org/working-groups/data-citation-wg.html

Making Dynamic Data Citeable

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned (aka history)

Researcher creates working-set via some interface:

- Access → **assign PID to QUERY**, enhanced with
  - **Time-stamping** for re-execution against versioned DB
  - **Re-writing** for normalization, unique-sort, mapping to history
  - **Hashing** result-set: verifying identity/correctness

  leading to landing page

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Prototype for CSV: http://datacitation.eu/

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve **original data** OR **current version** OR **changes**
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ubset of data
- Upon executing selection („download") user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve **original data** OR **current version** OR **changes**
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

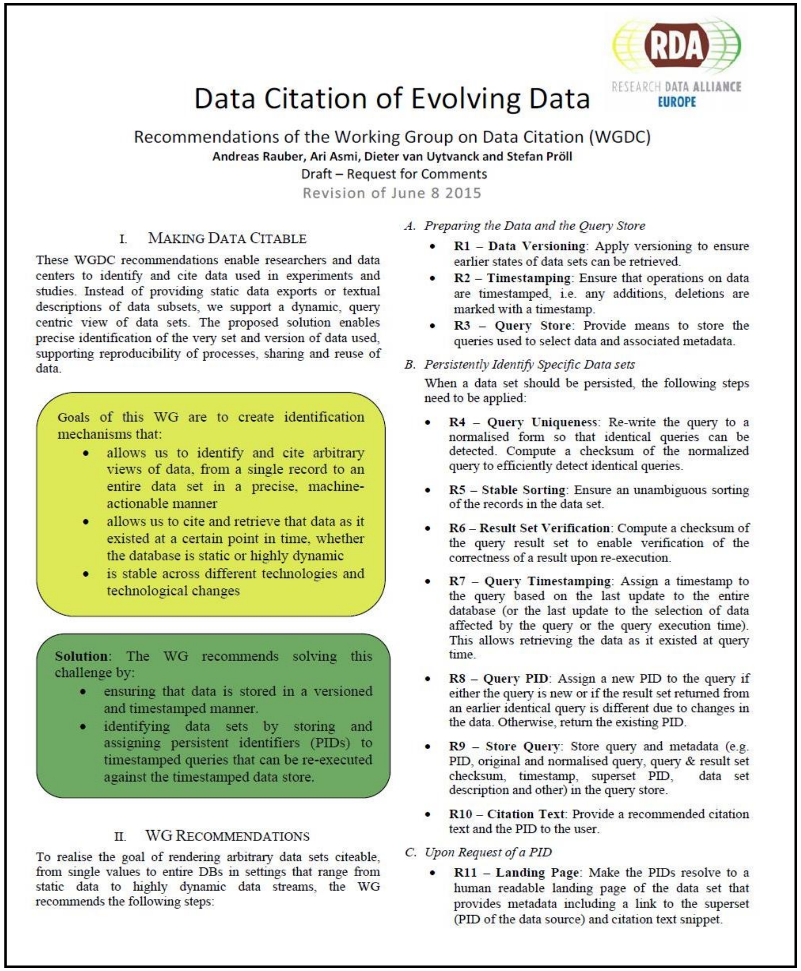
- ubset of data
- Upon executing selection („download") user gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citati n text (e.g. bTeX)

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

- PID resolves to land g page
  - Provides detailed metadata, link parent data set, subset,…
  - Option to retrieve **original data** OR **current version** OR **changes**
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned

Data Citation – Recommendations

- 2-page flyer,
  more extensive doc to follow
- **14 Recommendations**
- Grouped into **4 phases**:
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Upon request of a PID
  - Upon modifications to the data infrastructure
- History
  - First presented March 30 2015
  - Major revision after workshop April 20/21
  - Series of webinars
    (next: June 24, 18:00 CEST)



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Draft – Request for Comments

Revision of June 8 2015

I. Making Data Citable

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very set and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

II. WG Recommendations

To realise the goal of rendering arbitrary data sets citeable, from single values to entire DBs in settings that range from static data to highly dynamic data streams, the WG recommends the following steps:

A. Preparing the Data and the Query Store

- **R1 – Data Versioning**: Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping**: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store**: Provide means to store the queries used to select data and associated metadata.

B. Persistently Identify Specific Data sets

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness**: Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.
- **R5 – Stable Sorting**: Ensure an unambiguous sorting of the records in the data set.
- **R6 – Result Set Verification**: Compute a checksum of the query result set to enable verification of the correctness of a result upon re-execution.
- **R7 – Query Timestamping**: Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at query time.
- **R8 – Query PID**: Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID.
- **R9 – Store Query**: Store query and metadata (e.g. PID, original and normalised query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store.
- **R10 – Citation Text**: Provide a recommended citation text and the PID to the user.

C. Upon Request of a PID

- **R11 – Landing Page**: Make the PIDs resolve to a human readable landing page of the data set that provides metadata including a link to the superset (PID of the data source) and citation text snippet.

Summary

- Trustworthy and efficient e-Science
- Need to move beyond preserving data
- Need to move beyond the focus on description
- Process Management Plans (PMPs)
- Preservation (and verification)
- Support for citing arbitrary subsets of dynamic data
- Data and process re-use as basis for data driven science
  - evidence
  - investment
  - efficiency

Done!?

Acknowledgements

- Johannes Binder
- Rudolf Mayer
- Tomasz Miksa
- Stefan Pröll
- Stephan Strodl
- Marco Unterberger

- TIMBUS
- SBA: Secure Business Austria
- RDA: Research Data Alliance

TIMBUS
TIMELESS BUSINESS

secure
sba-research.org

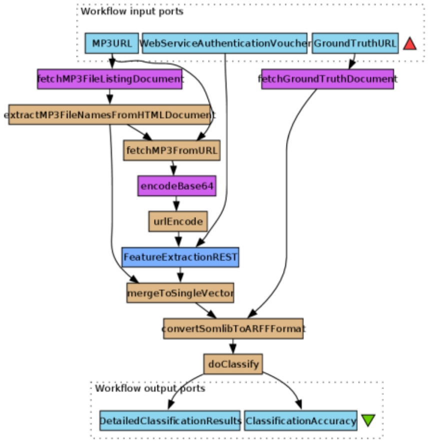
RDA
RESEARCH DATA ALLIANCE

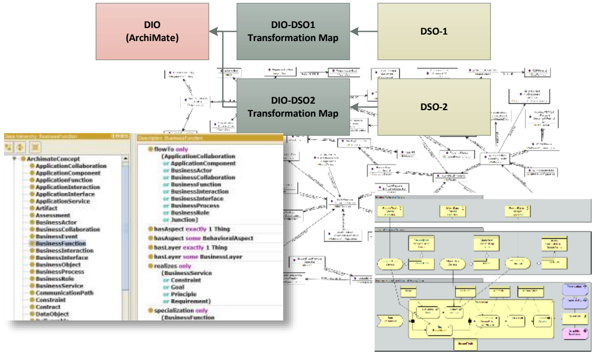
ifs FACULTY OF !NFORMATICS

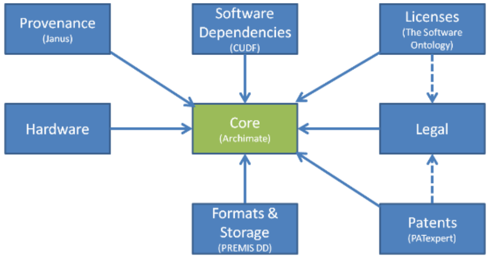
References

- Tomasz Miksa, Rudolf Mayer, and Andreas Rauber. Ensuring sustainability of web services dependent processes. International Journal of Computational Science and Engineering (IJCSE). 2015 Vol.10, No.1/2, pp.70 – 81
- Kevin R. Page and Raul Palma and Piotr Holubowicz and Graham Klyne and Stian Soiland-Reyes and Daniel Garijo and Khalid Belhajjame and Rudolf Mayer, "Research objects for audio processing: Capturing semantics for reproducibility," in 53rd AES International Conference on Semantic Audio (AES 2014), 2014.
- Tomasz Miksa and Rudolf Mayer and Stephan Strodl and Andreas Rauber and Ricardo Vieira and Goncalo Antunes, "Risk driven selection of preservation activities for increasing sustainability of open source systems and workflows," 11th International Conference on Digital Preservation (iPres 2014), 2014.
- Rudolf Mayer and Tomasz Miksa and Andreas Rauber, "Ontologies for describing the context of scientific experiment processes," in 10th Intl. Conference on e-Science, 2014.
- Tomasz Miksa and Stefan Proell and Rudolf Mayer and Stephan Strodl and Ricardo Vieira and Jose Barateiro and Andreas Rauber, "Framework for verification of preserved and redeployed processes," in 10th International Conference on Preservation of Digital Objects (IPRES2013), 2013.
- Tomasz Miksa, Stephan Strodl and Andreas Rauber, Process Management Plans. International Journal of Digital Curation, Vol 9, No 1 (2014),pp. 83-97. DOI:10.2218/ijdc.v9i1.303
- Rudolf Mayer and Mark Guttenbrunner and Andreas Rauber, "Evaluation of preserved scientific processes," in 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), 2013.

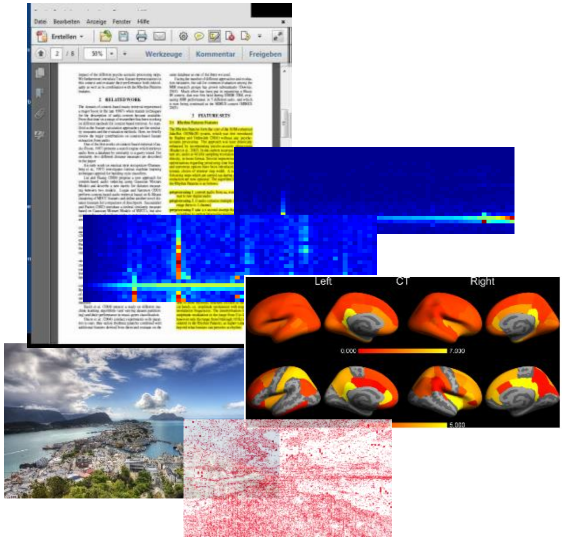
Thank you!

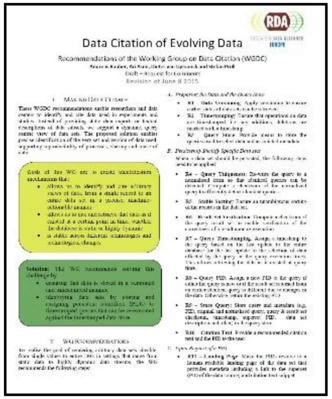






http://www.ifs.tuwien.ac.at/imp





Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Draft – Request for Comments

Revision of June 8 2015

RESEARCH DATA ALLIANCE

United we stand

DC1

Data Citation Principles

TIMBUS

TIMELESS BUSINESS

ifs FACULTY OF !NFORMATICS