TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

The MASi Research Data Management Service

Richard Grunzke, Volker Hartmann, Thomas Jejkal, Ajinkya Prabhune, Rainer Stotzka, Alexander Hoffmann, Sonja Herres-Pawlis, Aline Deicke, Torsten Schrade, Helen Kollai, Hendrik Herold, Gotthard Meinel, Wolfgang E. Nagel

richard.grunzke@tu-dresden.de

ZIH
Center for Information Services & High Performance Computing

Data Challenge

- Scientific knowledge gain increasingly based on data
- Amount of data and number of files continuously growing
- Limits in current approaches
- Towards FAIR principles [datafairport.org]
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**e-usable
- Novel methods needed

Metadata as Solution

- Data Management via metadata (“Data about data”)

  → Access data via content instead of name/location

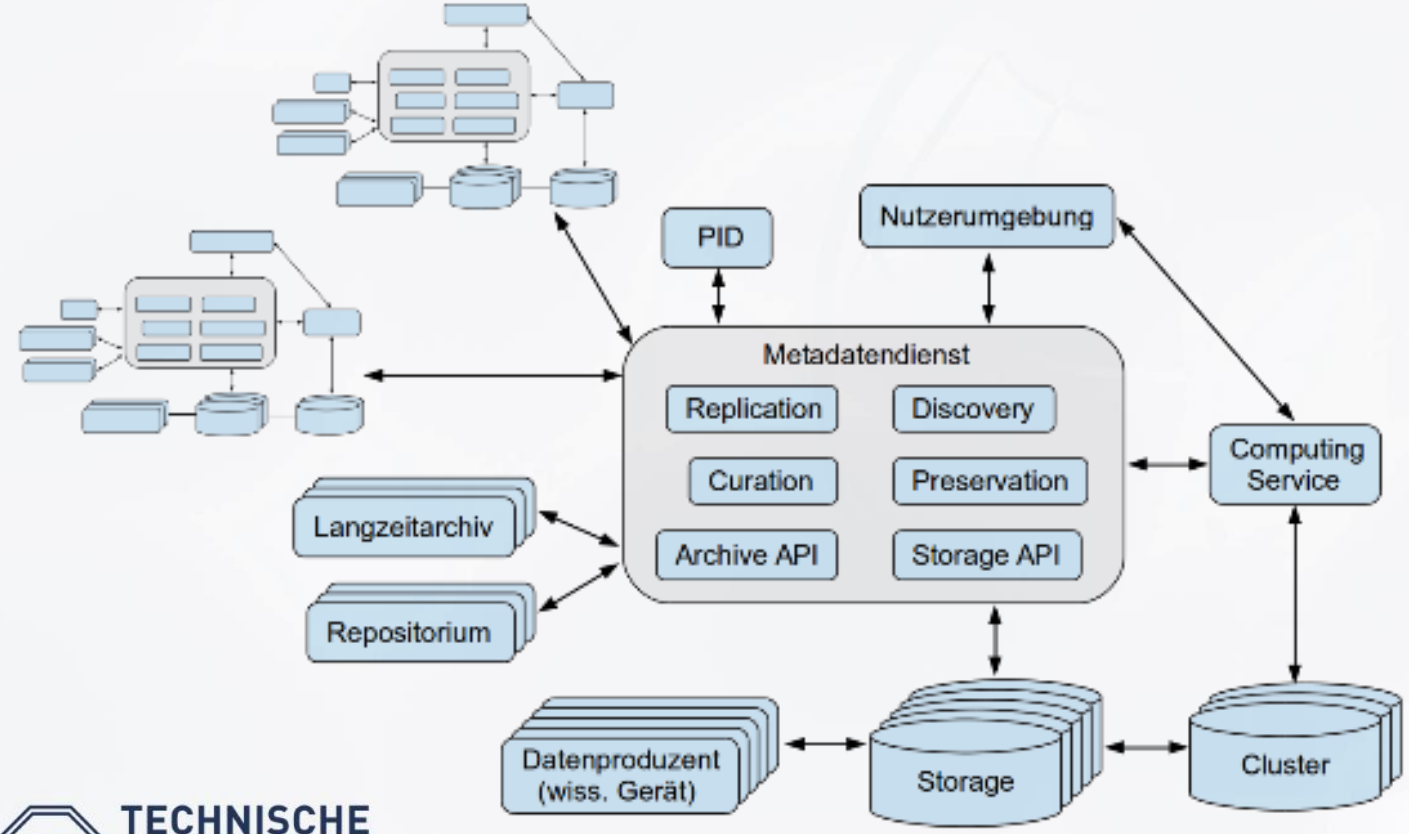
  → Find specific files in data set of millions of files
- Automatically extract, store and index metadata
- Metadata significantly improves usefulness of data

MASi – Project Facts

- Metadata Management for Applied Sciences
- Build-up metadata-driven research data repository for highly diverse communities
- Computer science partners:
  - ZIH at Technische Universität Dresden (lead)
  - IPE at Karlsruhe Institute of Technology
- Community partners
  - Geography - IÖR Dresden
  - Chemistry - RWTH Aachen University
  - Digital humanities - TU Darmstadt
    - Academy of Sciences and Literature Mainz
- DFG funding, 03/2015 – 02/2018

MASi – Overarching Goals

- Use and extend KIT DM framework - no development from scratch
  - Generic API – support arbitrary metadata models
  - Generic GUI – easily adaptable to community needs



MASi – Overarching Goals

- Create best practice integration guide
- Support further communities
- Collaborate within RDA
- Work towards sustainable service offer

MASi – Background – KIT Data Manager

- KIT Data Manager (KIT DM)
  - A metadata-driven repository framework
- Core development at IPE/KIT (Group of Rainer Stotzka)
- Basis for MASi service
- Generic, customizable, open source
- First public version in 2014, now version 1.3
- Well documented & open to contributions
- Used and extended in several projects
- http://datamanager.kit.edu/

MASi – Background – KIT Data Manager

- Organized into high-level services
  - Data and metadata management
  - Sharing
  - Administration
  - Data transfer
  - Data workflows
- Java and REST APIs
- Far reaching automation capabilities

MASi – Generic Metadata API

- Metadata to cover whole data life cycle
- Single point of access for all data
- High extensibility via REST interface
- PIDs to uniquely reference data
- Support widely used standards

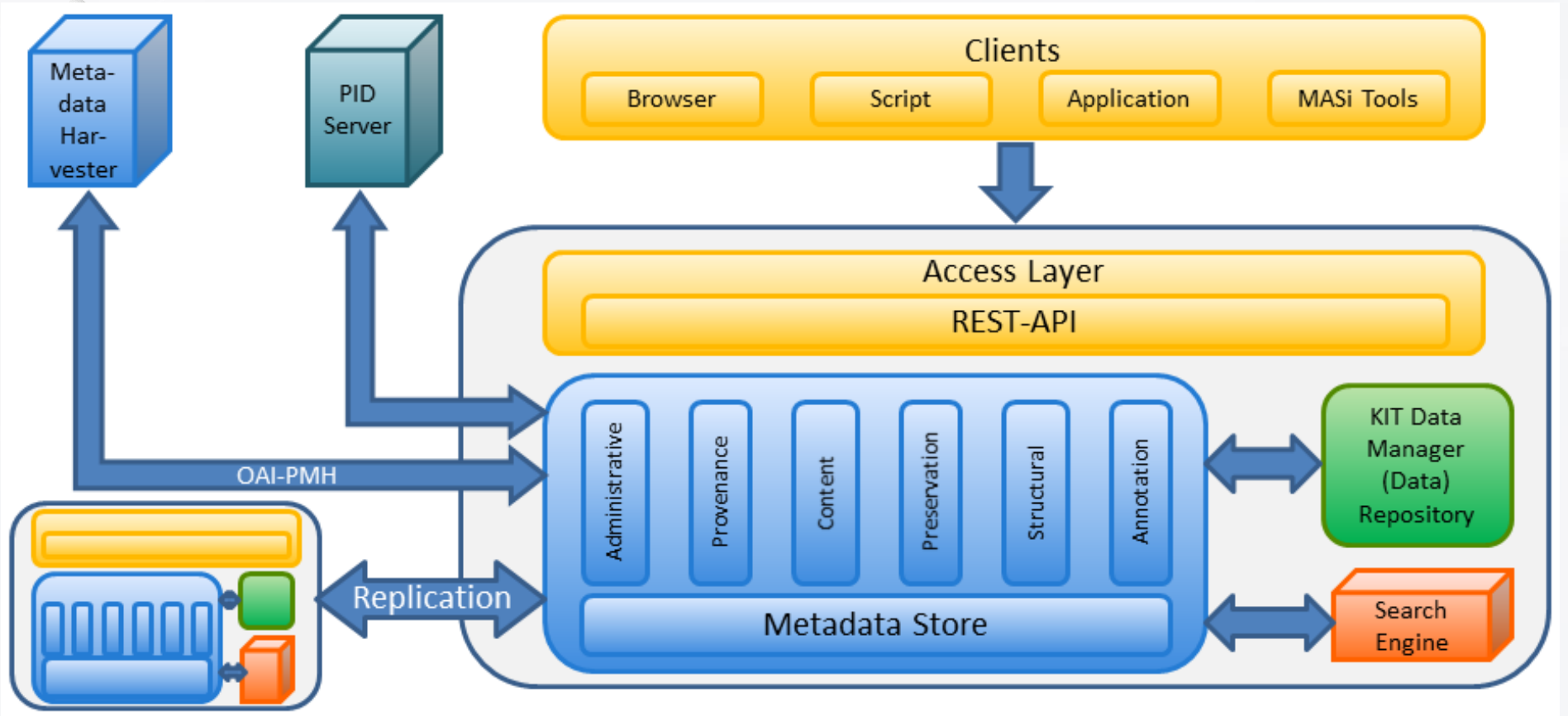
MASi – Generic Metadata API

- Metadata available as METS documents
  - “Metadata Encoding and Transmission Standard”
- METS defines various sections for different purposes
- Metadata in MASi split in packages similar to METS
  - Administrative, structural, content, provenance and annotation
- Appropriate combination of packages for specific community
- For each package possibly specialized database

MASi – Generic Metadata API

- Example for provenance package functionality
- Many workflow engines use separate workflow formats
- Many support Open Provenance Model (OPM) and ProvOne
- Graphs stored in graph database
  - Sophisticated queries such as “similar workflows”
- METS has pre-defined provenance section

MASi – Generic Metadata API



MASi – Generic GUI

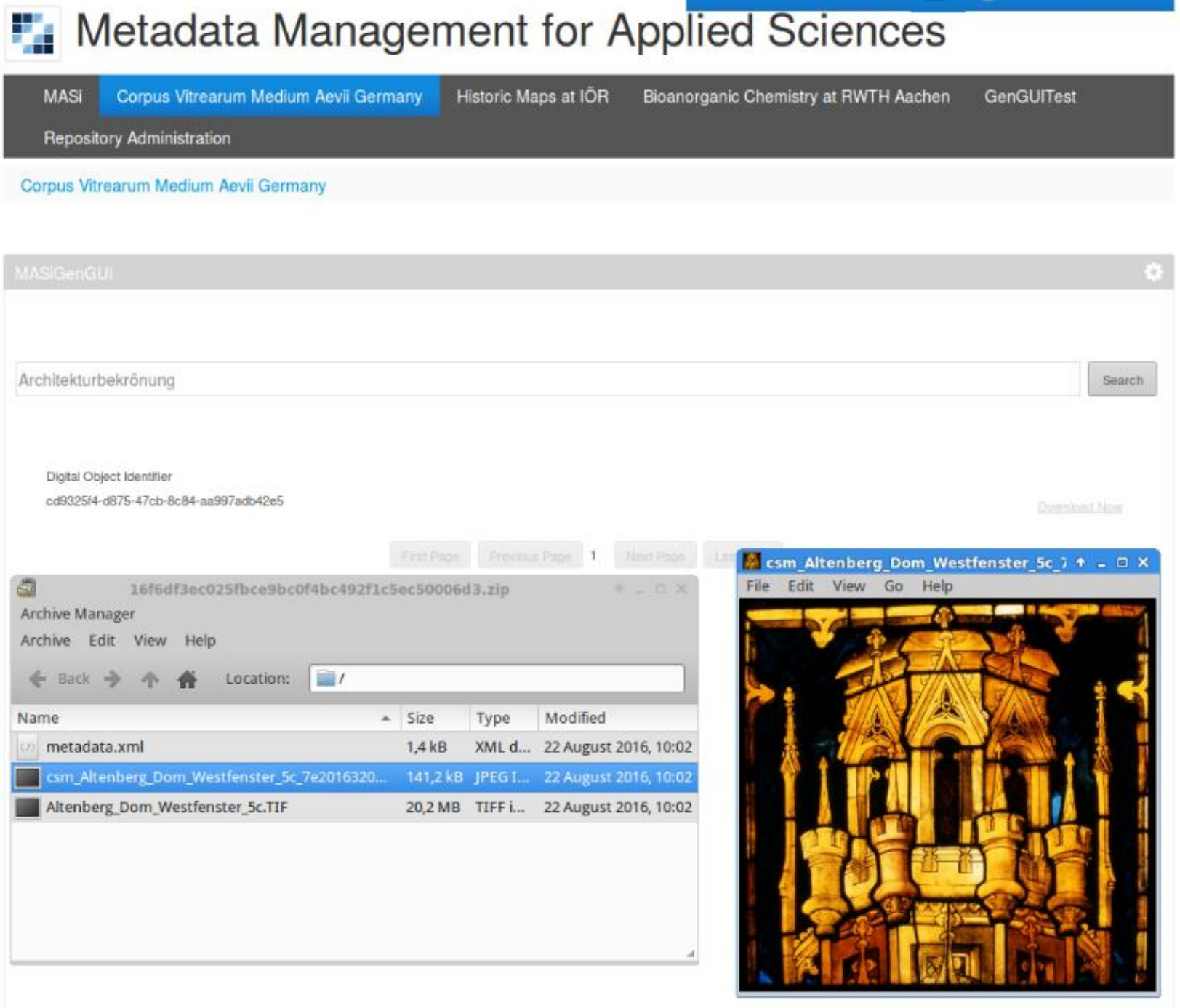
- Generic user interface to be specifically adapted
- Motivation to lower effort to adapt MASi to further use cases
- Time-saving for developers familiar with the technology
- Fundamentally enabling for developers unfamiliar with it
- Convenient access for users, specific to their use case

MASi – Generic GUI

- Built on basis of Liferay portal framework
- Development of plugin for KIT DM integration
  - Consistent user management between Liferay and KIT DM
- Liferay as ground truth
- Automatic syncing users and groups to MASi KIT DM instance
- MASi then transparently supports AAI methods that Liferay does

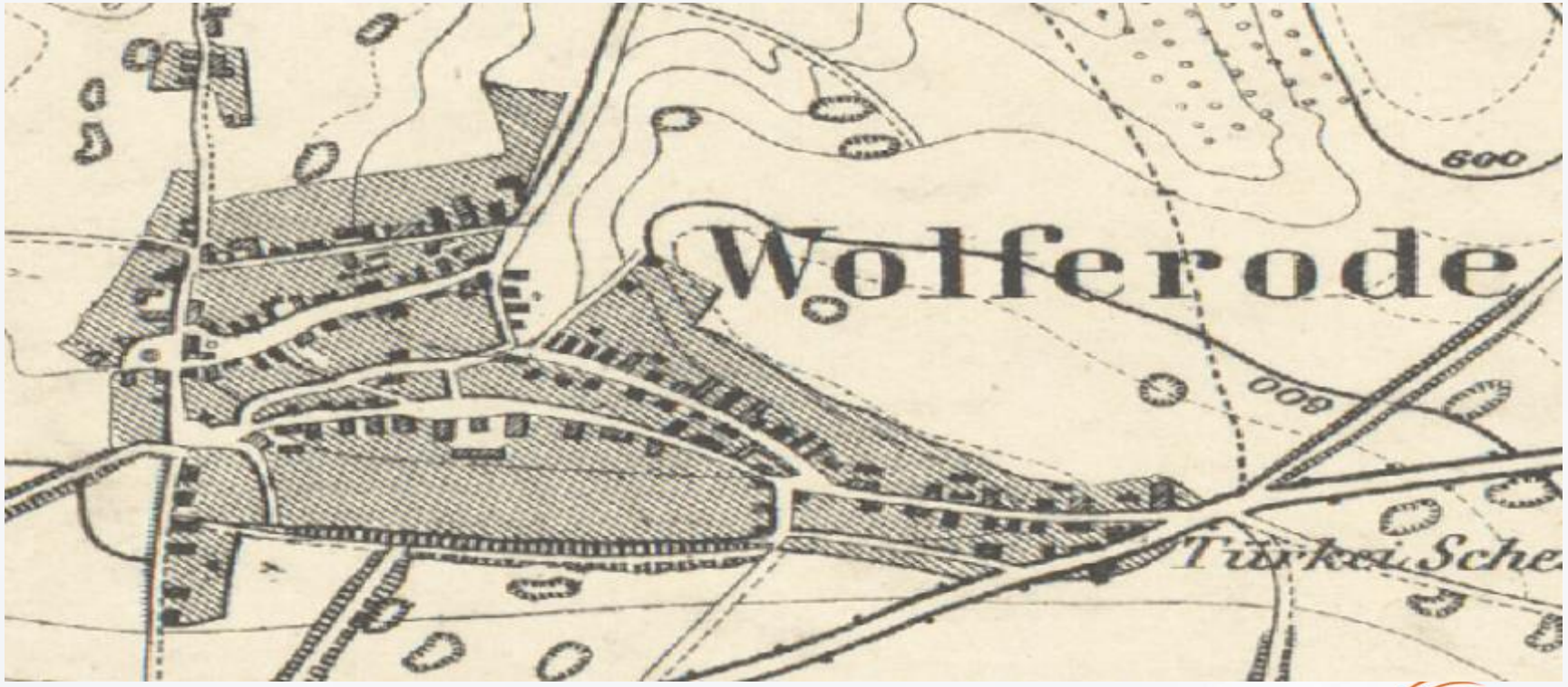
MASi – Generic GUI

- Development of generic portlet based on Vaadin library
- Continuous extension of functionality
  - For increasingly advanced generic capabilities
- Enables quick community adaptations
- Extensive documentation
- Open source and upstreaming to KIT DM framework


Admin
My Sites
0
Richard Grunzke
Metadata Management for Applied Sciences
MASi
Corpus Vitrearum Medium Aevii Germany
Historic Maps at IÖR
Bioanorganic Chemistry at RWTH Aachen
GenGUITest
Repository Administration
Corpus Vitrearum Medium Aevii Germany
MASiGenGUI
Architekturbekrönung
Search
Digital Object Identifier
cd9325f4-d875-47cb-8c84-aa997adb42e5
Download Now
First Page
Previous Page
1
Next Page
16f6df3ec025fbce9bc0f4bc492f1c5ec50006d3.zip
Archive Manager
Archive Edit View Help
Back
Location:
/
Name
Size
Type
Modified
metadata.xml
1,4 kB
XML d...
22 August 2016, 10:02
csm_Altenberg_Dom_Westfenster_5c_7e2016320...
141,2 kB
JPEG I...
22 August 2016, 10:02
Altenberg_Dom_Westfenster_5c.TIF
20,2 MB
TIFF i...
22 August 2016, 10:02
csm_Altenberg_Dom_Westfenster_5c_7
File Edit View Go Help

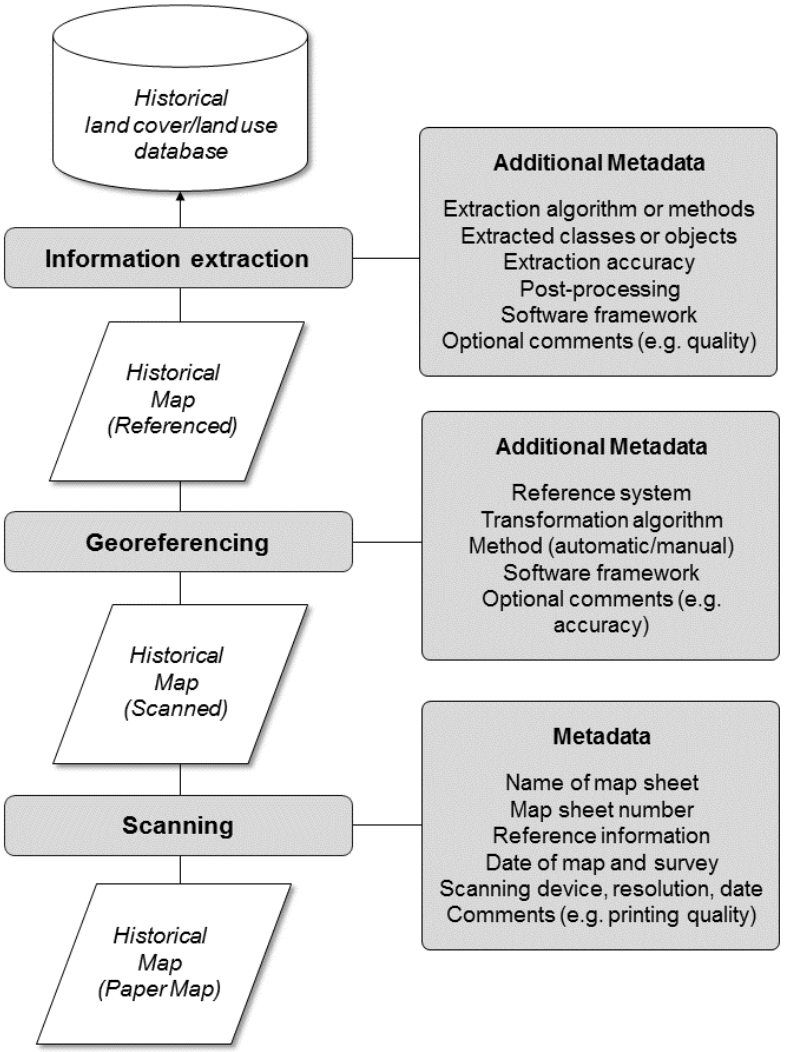
MASi – Use Cases – Historical Maps

- Historical maps for reconstructing land use changes
- Large scale spatial analyses and change detection
- Image analysis and pattern recognition algorithms on map scans



MASi – Use Cases – Historical Maps

- Metadata essential for change detection and interpretation
- Three major components
  - Scanning of paper maps
  - Georeferencing of scanned maps
  - Information extraction
- Automatic metadata extraction



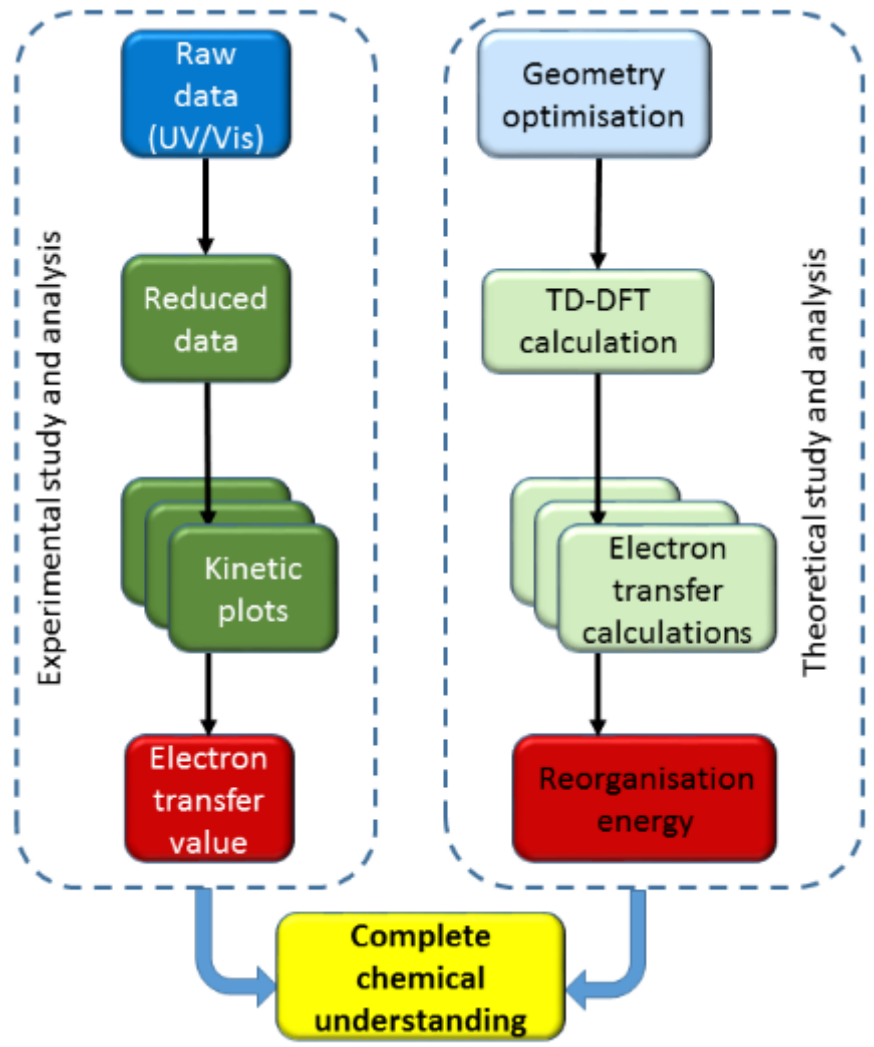
MASi – Use Cases – Spectroscopy in Chemistry

- Multitude of information by instruments
  - UV/Vis, IR, Raman, EPR and XAS spectroscopy
- Monitoring every 1.5 ms via instrument

  → Large amount of raw data
- Data currently manually reduced by choosing suitable characteristics
- Combination with theoretical simulation data highly fruitful

MASi – Use Cases – Spectroscopy in Chemistry

- Different levels of data production
- Metadata annotation in all steps
- Manually attaching metadata that is only known to users
- Handling of
  - Reduced data
  - Metadata for theoretical data



MASi – Use Cases – Church Windows

- Corpus Vitrearum Medii Aevi (CVMA) Germany
- Long-term research project for medieval stained glass windows
- Study each window's history such as iconography and religious context
- Curates image archive with set of XMP metadata for each TIFF file

MASi – Use Cases – Church Windows

- MASi to open up CVMA image archive to further interested parties
- E.g. providers of cultural heritage photography
- Use of upcoming OAI-PMH interface
- Automatic functionality
  - Matching of metadata records with other cultural repositories
  - Annotating of CVMA images with these records

MASi – Sustainability

- Further use cases
- Collaborations
- Towards service offer
- Further technology improvements

MASi – Sustainability – Further Use Cases

- Basis of and extensions in further use cases (up to 24 user groups each)
- DFG Collaborative Research Centre 940 - Volition and Cognitive Control
  - Psychology, neuroimaging
- DFG Collaborative Research Centre 980 - Episteme in Motion
  - History
- H2020 Nanoscience Foundries & Fine Analysis (NFFA)
  - Nanoscience
- DFG Generic Research Data Infrastructure (GeRDI)
  - Ocean, environment and life sciences, economy, digital humanities
- Further proposals underway

MASi – Sustainability – Collaborations

- Contribute to RDA and evaluate integration of RDA recommendations
- ScaDS Dresden/Leipzig Big Data competence center
  - MPI-CBG, UFZ, UKD
- Nanoscopy reference archive with KIP (Heidelberg) and IMB (Mainz)

ScaDS DRESDEN LEIPZIG

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

MASi – Sustainability – Towards service Offer

- General service offer for arbitrary user communities
- Planning phase
- Aim at production mode
  - Support, bug fixes, new features, scheduled downtimes
- Utilization of IT infrastructure at ZIH / TU Dresden

MASi – Sustainability – What is missing?

- Better support arbitrary communities
  - Schema hierarchies
  - Controlled vocabularies
  - Extended automatic validation
  - Support definition of practical policies

MASi – Summary

- MASi enables
  - Metadata-driven management of large and complex data
  - Quick integration of highly diverse use cases
- Uses and extends open source framework KIT DM
- Work towards production-ready and sustainable service offer
- Need to even better support arbitrary community use cases

Thank you