MAX PLANCK
digital library

eInfrastructures: Simple is beautiful

Laurent Romary
Max Planck Digital Library

GoeGrid

Göttingen, 13 September 2007

Background

- Many initiatives to provide research infrastructures at national and EU level
- What do they or will they offer?
- Do we match the scientists' expectations?
- Access, standards and community building.

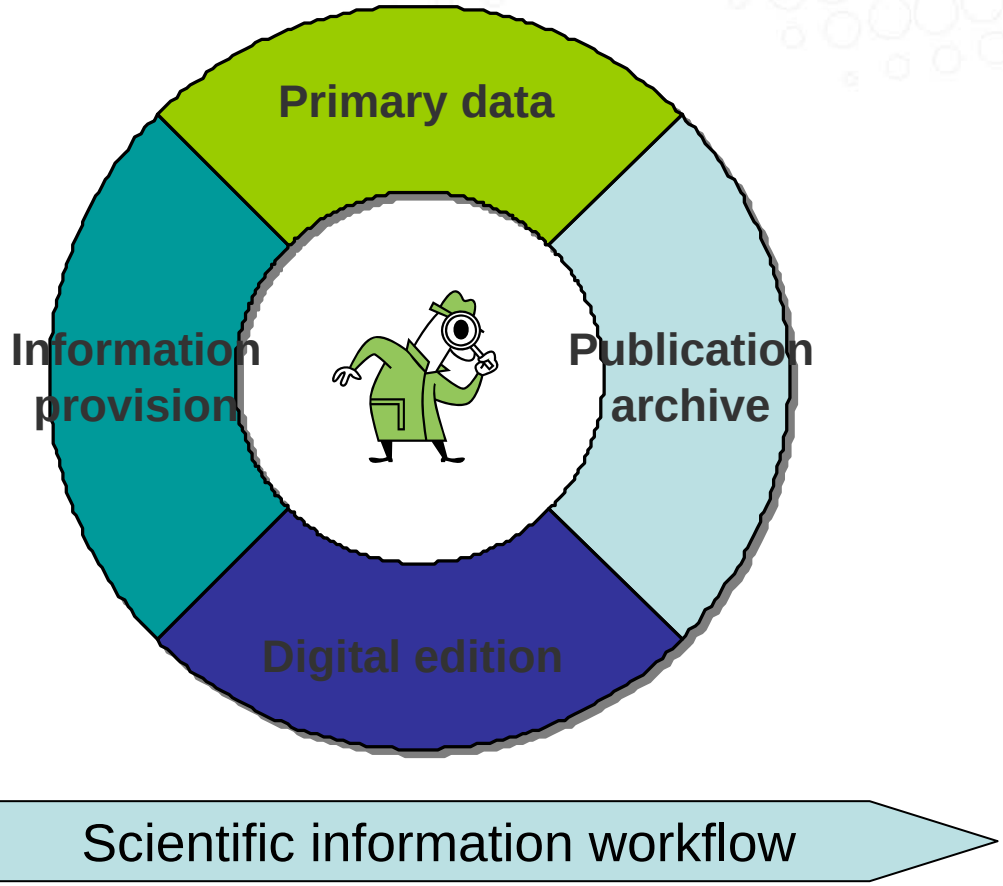
…with a view on the humanities

Why do we need eInfrastructures?

- The scientist's ecology

- Dealing with digital sources

The Scientist's (digital) ecology



Working with primary (digital) sources in the humanities

Metadata

Annotation

Transcription

Metadata

Translation

Metadata

Annotation

Research Infrastructures

- RIs in general: permanent and physical
- RIs for the natural sciences
  - ice breakers for polar research, satellites, telescopes, particle accelerators, laboratories
- RIs for the humanities?
  - Cultural heritage in all forms is the main source of humanities research
  - Libraries and archives are the traditional “laboratories” for the humanities
- In the digital age, essential for innovative humanities research is:
  - Access to digitised heritage data (data bases, text corpora, speech, image collections, etc.)
  - Tools to process this information

European Strategy Forum for Research Infrastructures

Digital Research Infrastructure for the Arts and Humanities
DARIAH

The role of the European Strategy Forum on Research Infrastructures (ESFRI) is to support a coherent approach to policy-making on research infrastructures in Europe, and to act as an incubator for international negotiations about concrete initiatives.

Goals of DARIAH

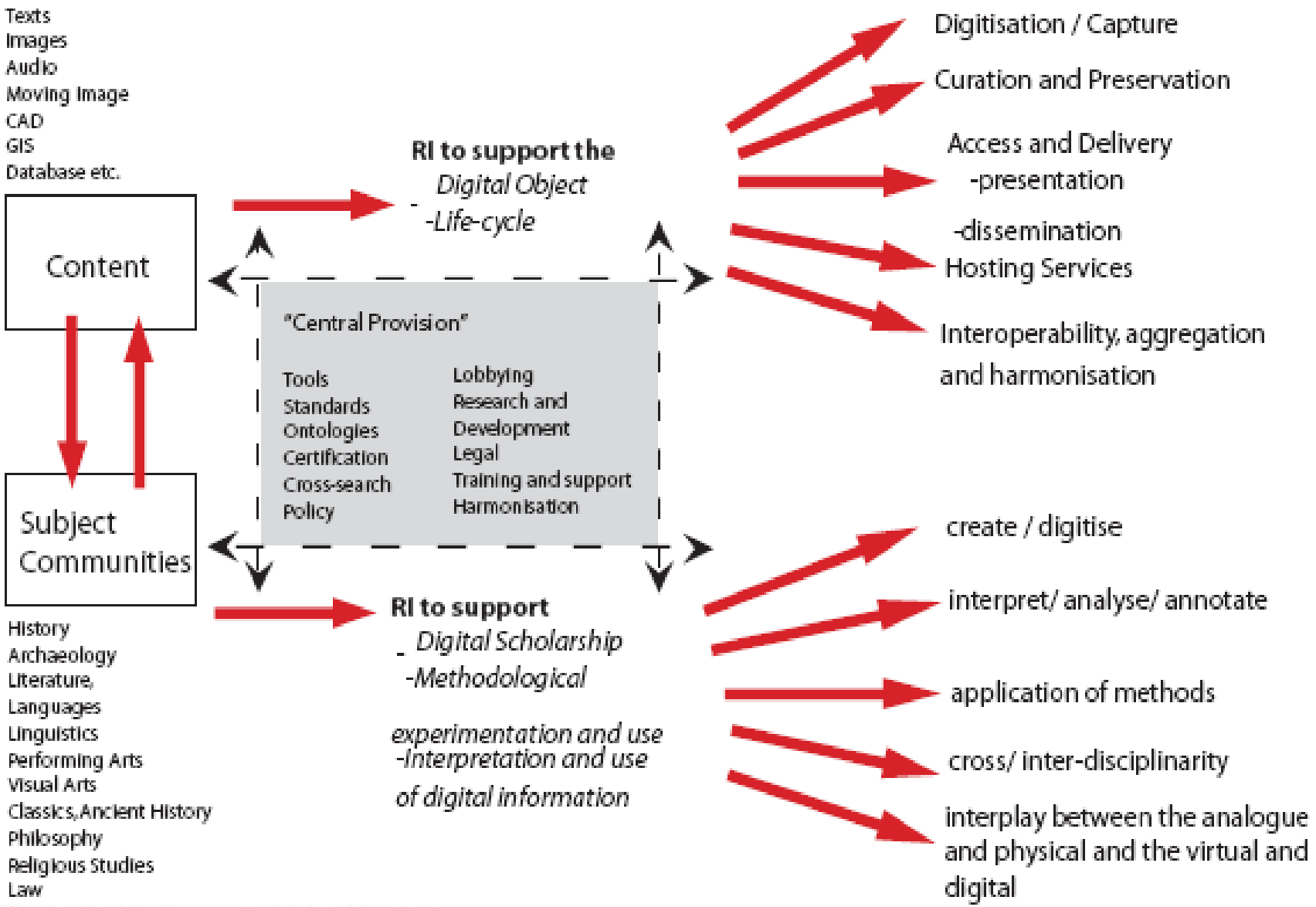
- Providing a coordinated infrastructure that would act as both a catalyst and support for the development of national services and digitisation programmes aimed particularly at those European countries without such services and programmes. A key feature of the Infrastructure is the expansion of the RI to include many other European nations: **BUILDING NATIONAL CAPACITY**

- Providing a coordinated infrastructure that would act as a catalyst to bring together the different sectors involved in cultural heritage and humanities information management and access – education, memory and cultural heritage institutions and organisations, and the commercial sector – in order that they might work together for the benefit of both themselves and the research communities across Europe: **BUILDING AN INTEROPABLE INFRASTRUCTURE**

Core activities

- Digitise – Curate – Preserve
  - Standards development and promotion
  - Curation, preservation and digitisation services
  - Technology platforms
  - Legal services and advice
- Discover – Access – Deliver
  - Authentication and authorisation,
  - Harvesting, aggregating, hosting
  - User-friendly discovery, delivery and use
- Connect – Collaborate – Use
  - Supporting communities of practice
  - Facilitating new research practice
  - Tools and registries

Research Infrastructure for the Humanities and Arts



The Max-Planck Society

Max Planck Society in figures

- **80 Institutes**
  - **basic research**
  - **all subject areas**
  - **distributed organization**
- **Budget**
  - **1.3 bill. EUR (~1.6 bill. US-$)**
- **12,000 employees**
  - **3,500 scientists**
  - **8,500 support staff**
- **9,100 annual visiting scholars**

The Max Planck Digital Library (MPDL)

- The newly created structure dedicated to scientific information within the Max Planck Society

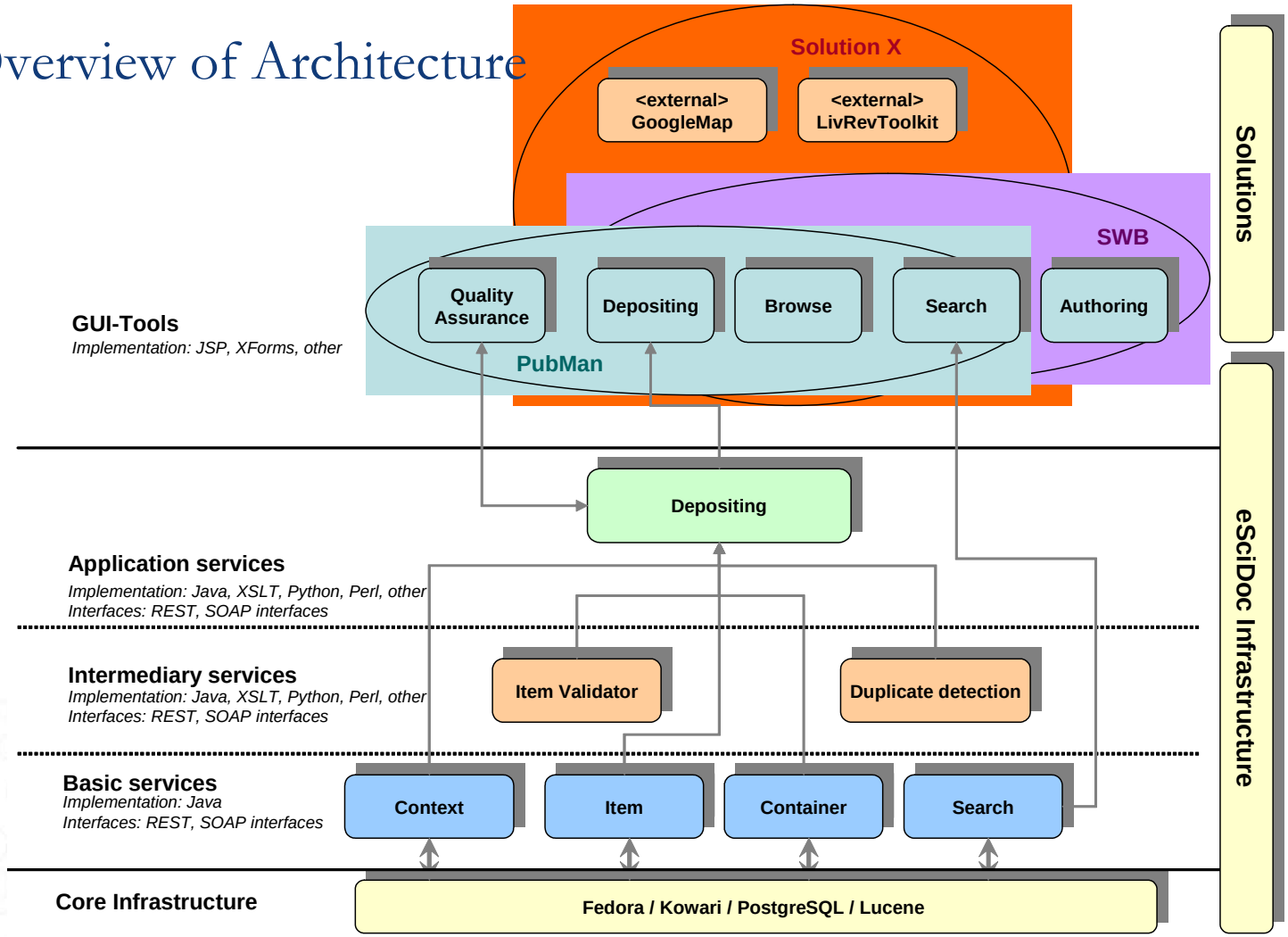
- Information sources
  - Journal and database subscriptions
  - Primary data and publications from researchers/institutes
  - Digital edition support
- Actors
  - Scientists
  - Librarians, IT support
  - Publishers, scholarly associations
  - Other institutions in Germany and beyond…

The eSciDoc project

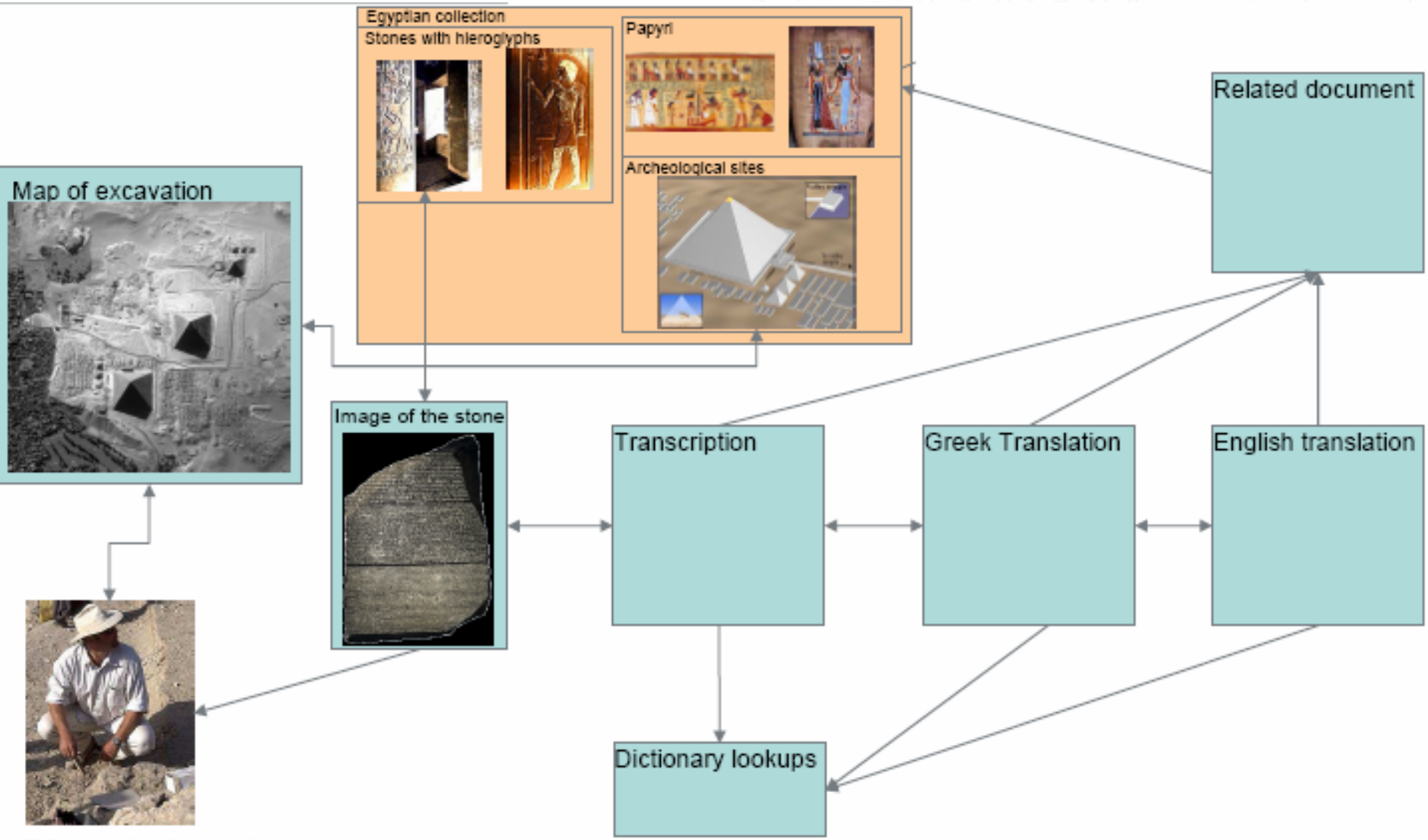
- Overview
  - Joint project of the Max Planck Society and FIZ Karlsruhe
  - Five-year grant (2004 – 2009) from the German Federal Ministry of Education and Research

- Target
  - a platform for flexible, open and persistent access to research results and materials
  - development of specialized solutions on top of a generic infrastructure addressing specific needs and requirements by Max-Planck Institutes… and beyond
- eSciDoc as a strategic project for the MPG
  - The basis for our future digital resource management activities

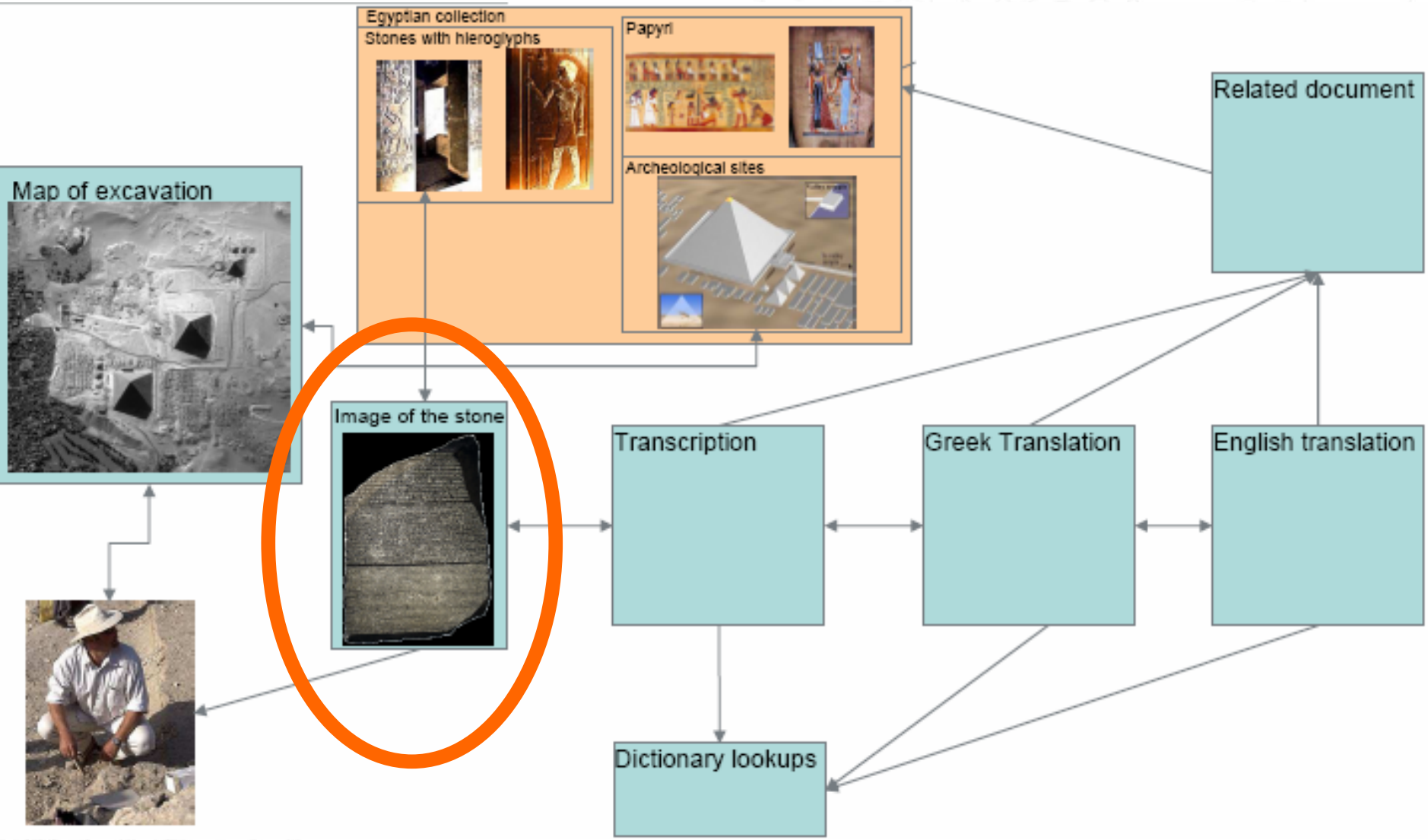
Overview of Architecture



SWB Modularization: Initial Context



SWB Modularization: Dealing with Images



Dealing with Images and Metadata (Solution Example 1)

- Max-Planck-Institut für Bildungsforschung

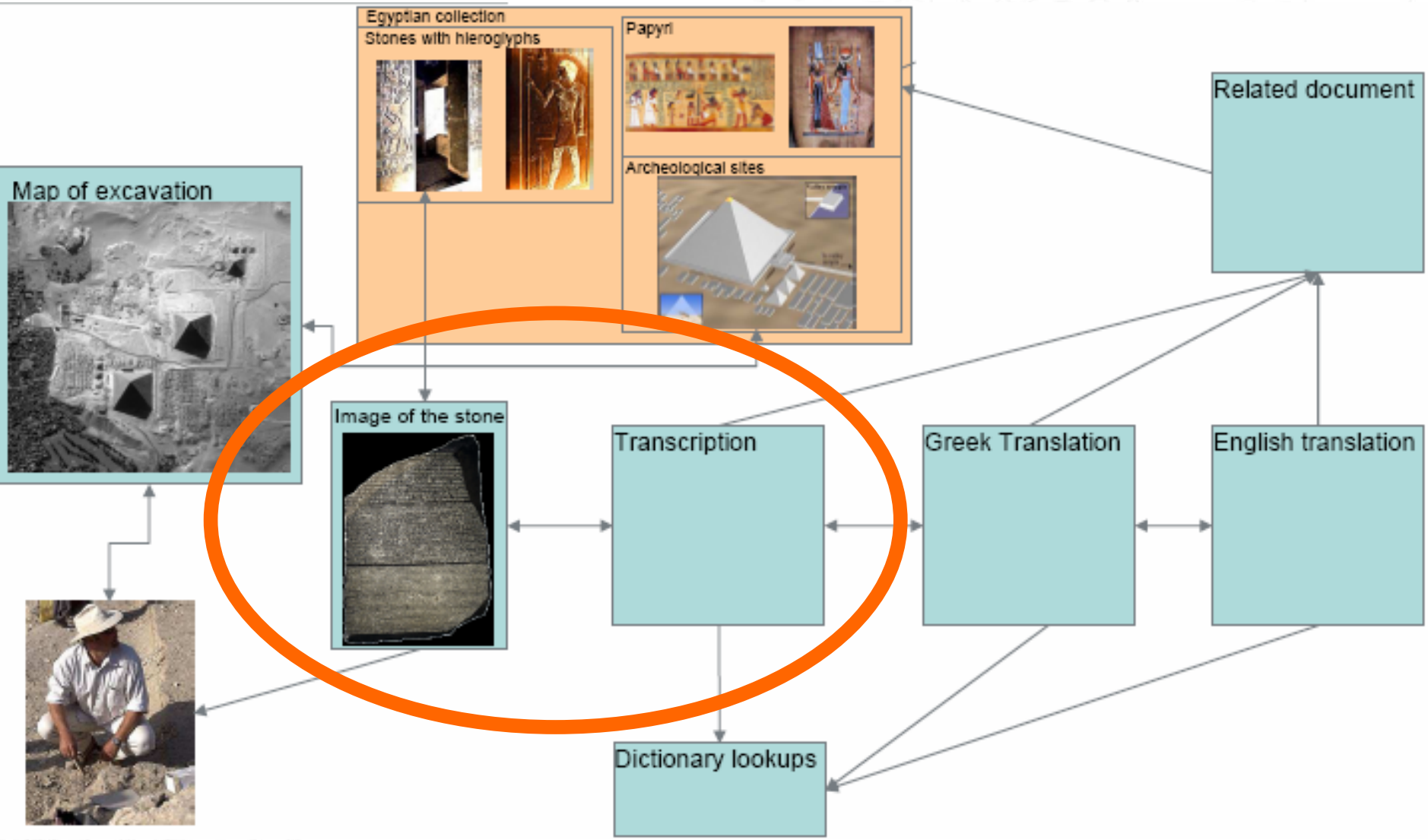
- Requirement to deal with Images
- Images of “Human Faces”
- Institute local collection
- Discovery and Retrieval of Images
- Metadata searches and image displays

QuickTime™ et un
décompresseur TIFF (LZW)
sont requis pour visionner cette image.

- First Assumption:
  - One MD-record for Images
  - One MD-record for Face specific Data

- Further needs: image annotation (details)

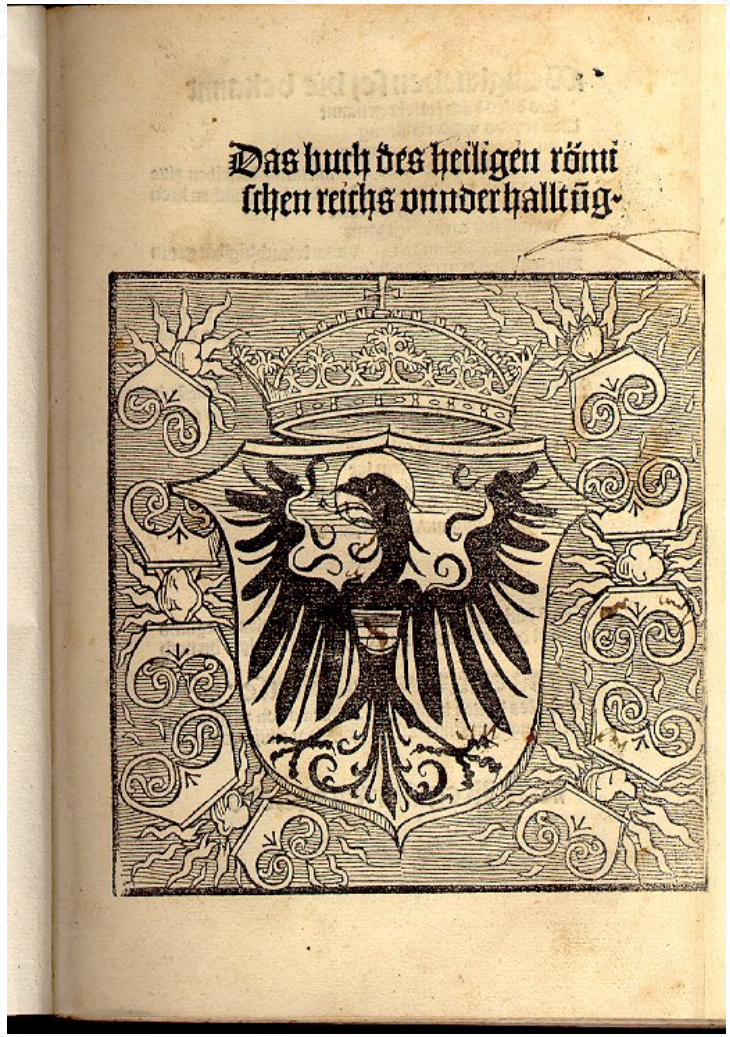
SWB Modularization: Images and Transcriptions



Dealing with Images and Transcriptions (Solution Example 2)

- Max-Planck-Institut für europäische Rechtsgeschichte

- Requirement to deal with digitized textual sources
- Legal documents (17th-19th)
- Content:
  - Precise bibliographical data
  - Page + full text transcription
  - Table of contents
- Further needs:
  - Text annotation



Some issues

- Any room left for new ideas?
- How can we accumulate expertise?
- Who will ensure the curation of data?
- Not to speak about open access…

MAX PLANCK
digital library

New ideas — living sources

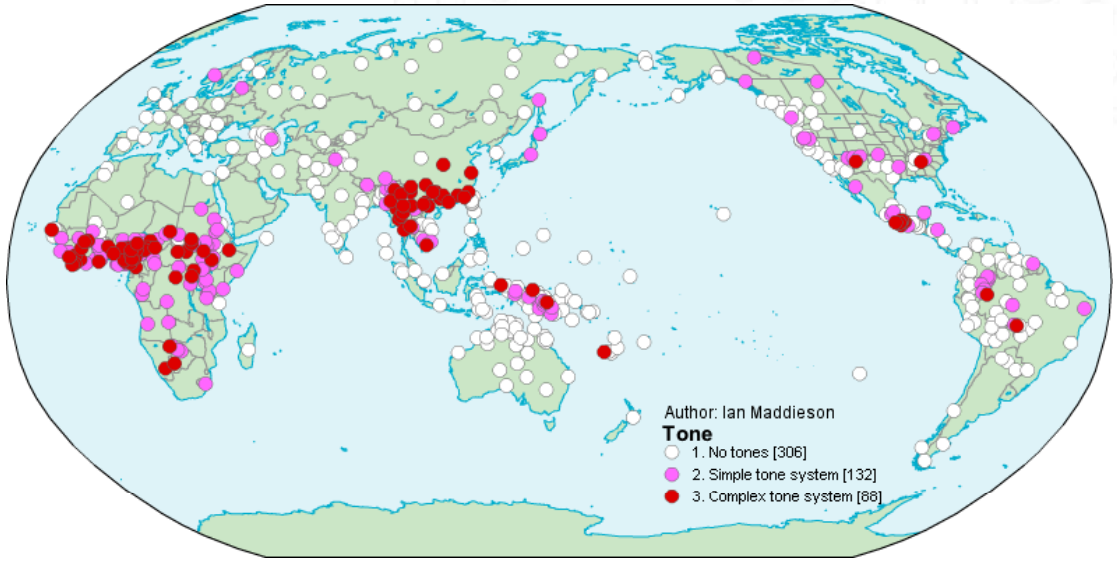
Describing observations

World Atlas of Language Structures (WALS)

•2560 different languages

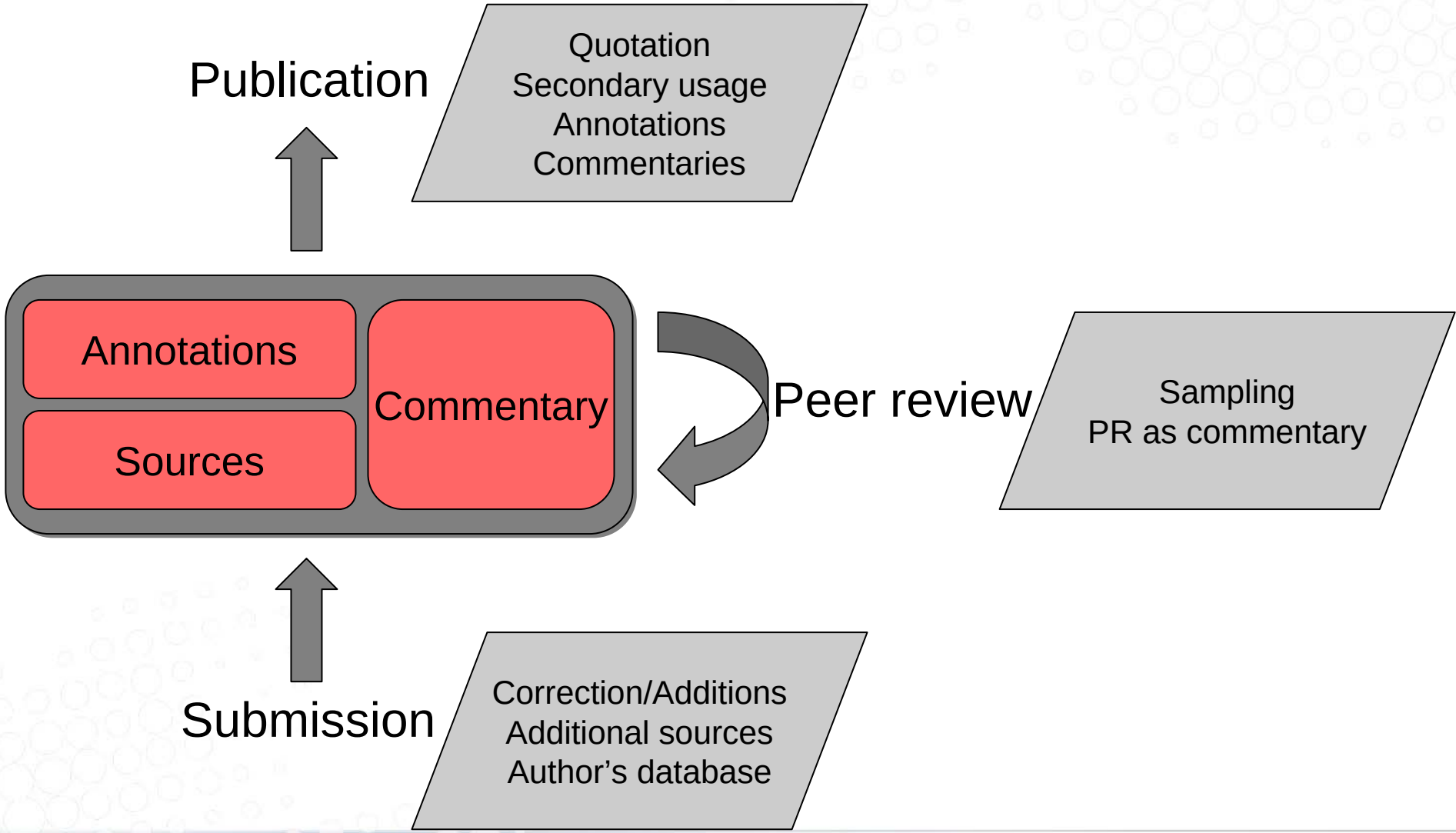
•142 features/maps/chapters

•Phonology, Morphology, Nominal Categories, Nominal Syntax, Verbal, Word Order, Simple, Complex Sentences, Lexicon, etc.



- Setting up a peer-reviewed environment
  - Submission of chapter + dataset
    - Conformance to good practice and scientific added values
  - Using WALS prestige
    - Getting academic credit for the data sets

Living sources



MAX PLANCK
digital library

Gathering expertise — Colab

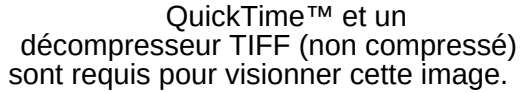
Standards and good practices

- An essential aspect of data preservation and reuse
  - Legibility of data
    - In space: sharing scientific sources with others
    - In time: pooling together the records of science
  - Generic standards (horizontal)
    - ISO 10646/Unicode, XML, etc.
  - Specific standards
    - ISO-IEC/JTC 1 (MPEG), ISO/TC 37 (ISO 639, TMF), TEI
  - E.g. TEI:
    - A wide range of documented elements for the encoding of textual data
    - A flexible architecture to select the elements adapted to one's needs

…potential complexity

MPDL CoLaboratory (MPDL CoLab)

- Platform for community building and knowledge exchange
- Aim:
  - improve exchange of explicit knwoledge and make tacit and individual know-how explicit
- Supports community-building processes
  - Connects people with similar fields of interest and goals
  - within the MPS: MPDL, librarians, scientists
  - Outside: underlying basis of our national and international collaborations
- Provide information about existing standards and best practices in the domain of supporting scientific life cycles
  - Ensuring long-term compatibility between local and centralized initiatives within the MPDL


QuickTime™ et un
décompresseur TIFF (non compressé)
sont requis pour visionner cette image.

Libraries, librarians: new scope, new roles

- Library as function
  - From information provision to information management
    - Identification of a “digital curator” profile: interface between scientists and scientific information
  - Local mirrors of central activities
    - We probably do need even more librarians…
- Library as a place
  - Core reference monographies
    - Complementarity with centralized archives
  - Local management of primary sources
    - Selection, digitization, access
    - Library as digital curation centres
  - Centre of gravity of scientific information (cf. Bibliothek2007)

Final words

- e-Infrastructures
  - We need them => which model fits which scientific community
- Communities
  - Sharing content and practices
- Central-decentral
  - A constant balance between contradictory forces

- Objective: simplicity for scientists