CLARIN

Common Language Resources and Technology Infrastructure

A European Research Infrastructure

Data Bases and Web Services for (a) Research Infrastructure(s)

Peter Wittenburg (MPG & CLARIN)

Peter Doorn (DANS & DARIAH)

Who is he?

- Technical Director at the Max-Planck-Institute for Psycholinguistics Nijmegen NL

  what happens in the brain while we are talking and listening -> data driven research

  ranges from typical humanities to biological methods (brain imaging with fMRI etc)

- member of the central IT board of the Max Planck Society

  as chair of an "archiving task force" I was responsible for a strategic decision

  - in 2004 Max Planck Society decided the following
    - the two CCs have to make a long-term archiving offer to any MPG researcher

      (my MPI's 50 TB are stored at 5 different locations for less than 10 k€ !!!)
    - data to be archived needs to be accompanied with proper metadata
    - anything beyond bit-stream preservation is left to the communities

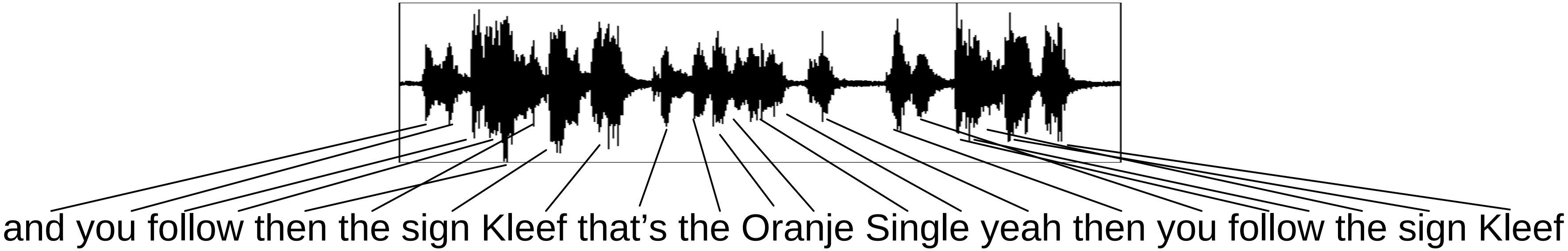
      (selection, MD set, format migration, terminology registration etc)
    - 50 years of "institutional backing" for all data assuming that MPG may exist for another 50 years, but perhaps not the CCs

- since 2008 responsible for the technical infrastructure in the CLARIN RI

Do we have a mission?

- CLARIN wants to create an integrated and interoperable domain of language resources and technology as an accessible service for all those researchers who work with language resources.

- we need to think of the small challenges - increase efficiency at the daily work of the researchers - and the big challenges
- small challenge: aligning speech and text via some stochastic machinery


and you follow then the sign Kleef that's the Oranje Single yeah then you follow the sign Kleef

- big challenge: improving speech recognition and/or machine translation for example
- no further PR: web-site, newsletter, Virtual Language Observatory

What kind of data?

- CLARIN and beyond such as DARIAH, CESSDA etc
  - typical time series data (speech, motion + eye tracking, EEG, fMRI etc)
  - audio/video recordings and tons of photos
  - text collections (corpora such as THE Dutch Spoken Corpus)
  - structured annotations on top of all these primary recordings in standoff fashion (different linguistic levels)
  - treebanks (syntax annotations of masses of texts)
  - structured lexica with multimedia extensions or links to fragments in archive
  - conceptual spaces ("kind" of ontologies), wordnets, etc
  - metadata descriptions as glue bundling and relating

- order of magnitudes: at MPI currently 50 TB of data, others certainly less
- what counts is not TB but the complexity within and between resources
- time series are comparatively simply structured
- AND: beyond UNICODE and XML there are no agreed standards

What will he talk about?

- already gave some background information

- repositories/archives and quality
- metadata
- virtual collections and integration
- workflow chains and interoperability
- (cost aspects)

LRT Situation

- about 150 members, i.e. institutions that have language resources and/or tools
- all is very fragmented, invisible and inaccessible
- CLARIN way:
  - cannot integrate 150 institutions - but need a backbone of service centres
  - need new types of service centres (“without own agenda, without burocracy”)
  - established criteria for such service centres

    (proper repository system, archiving strategy, quality assessment, MD, PID, part of a service provider federation, access APIs etc)
  - no requirement wrt repository system (iRods, FEDORA, D-Space, eScidoc, LAMUS, etc) - but we are asked to give advice and help
- about 30 institutions want to become such a centre
  - talked with all of them as a kind of assessment
  - almost all are busy with restructuring their holding !!!
  - almost all are talking with their national grid/CC/federation experts

Repositories/Archives

- task: store data and enable accessibility and enrichments in a way that when I have an identifier I will get exactly that resource I am expecting
- let’s not forget: research collections are “living entities”
- persistent identifiers, version control, authenticity checks are a MUST

  take care: we are speaking about millions of PIDs and add. functions

  this is not the DOI business model which is good for publications etc
- ESFRI document: Availability of data, Permanency, Quality, Rights of use, Interoperability (what does this imply?)
- wrt archiving (or long term preservation - most of the data for ever)
  - only few thought of this
  - only two institutions offer “open deposits” and have a long-term strategy
  - these two cannot take “all” (not a matter of terabytes)
  - we clearly miss a sustainable infrastructure with clear APIs

Quality

- increasingly important
- where do we talk about?
  - quality of data or quality of repositories/archives?
- quality of data
  - formal correctness - can check this if there is a schema
  - content correctness - only peer review system may work
  - but who has the time, who has the knowledge, who has the money
  - why not make it re-usable and let experts comment if they are interested
- quality of repositories/archive
  - they should establish rules about major aspects and make them visible
  - regular self-assessment such as Data Seal of Approval (DANS) to get certification much more useful than any OAIS based checks
  - rules should include formal correctness check, check on MD and association with PID (incl. authenticity information) at upload time
  - preservation strategy MUST be clear

Metadata

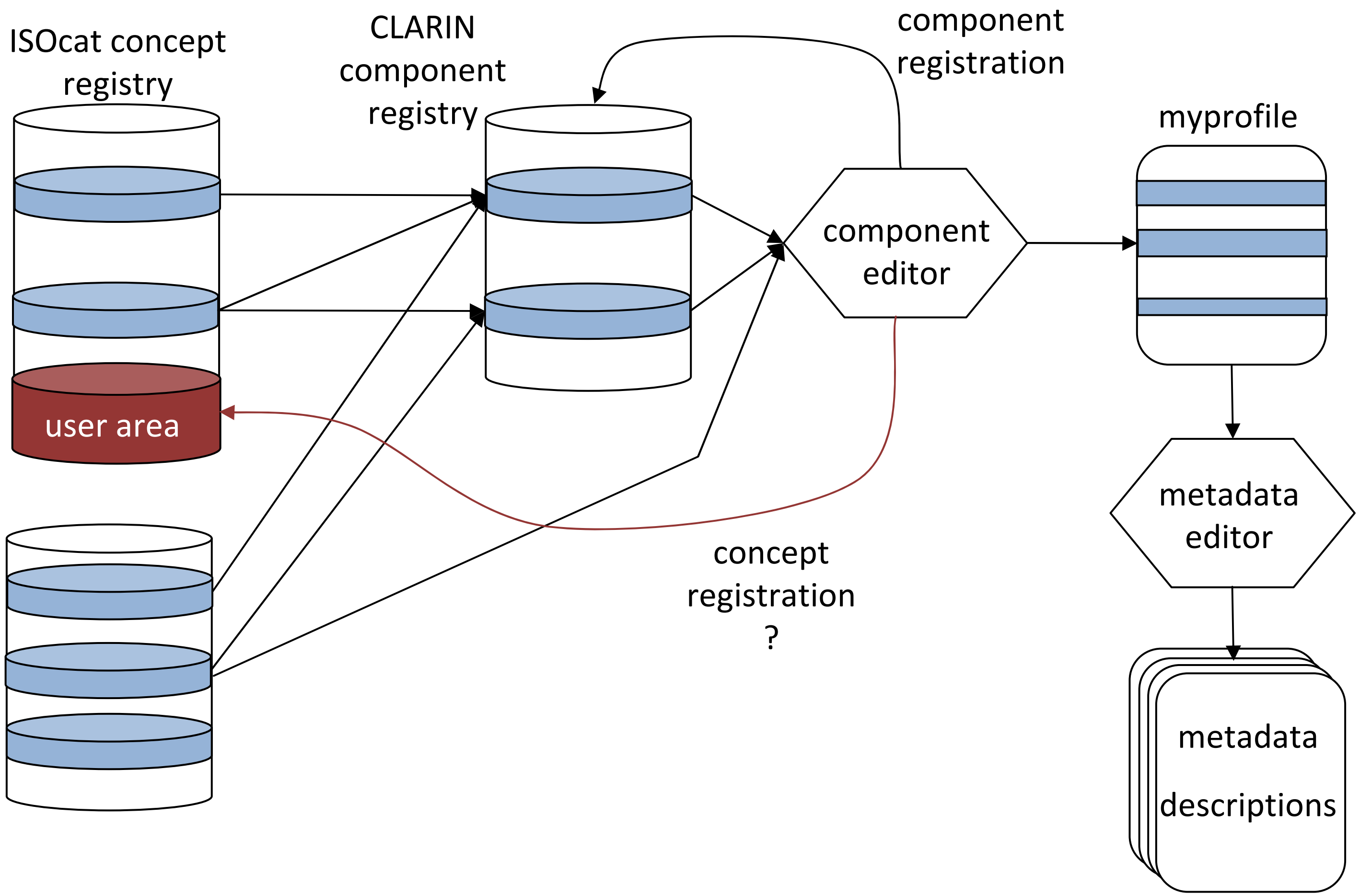
- about two decades of practical experience with metadata for electronic resources
- basically two approaches:
  - generic sets motivated by digital library experts (Dublin Core)
  - domain-specific sets worked out by domain experts (IMDI, LOM, VO, AAT, so many)
  - main differences:
    - MD is part of the research process (specific research questions etc)
    - need domain terminology, specific semantics mirroring the data types and the knowledge, flexible extension mechanisms etc
  - both are a fact and often gateways to Dublin Core for example are provided
- conclusions so far
  - the current coverage (IMDI, OLAC) is not sufficient
  - a single schema approach with embedded semantics is not sufficient
  - there are even sub-discipline differences and flexibility requirements are enormous
  - separate “concept” (data category) definitions to make them re-usable
  - allow users to create their own schemas by referring to registered categories
  - rely on PIDs for all the references

CLARIN MD State

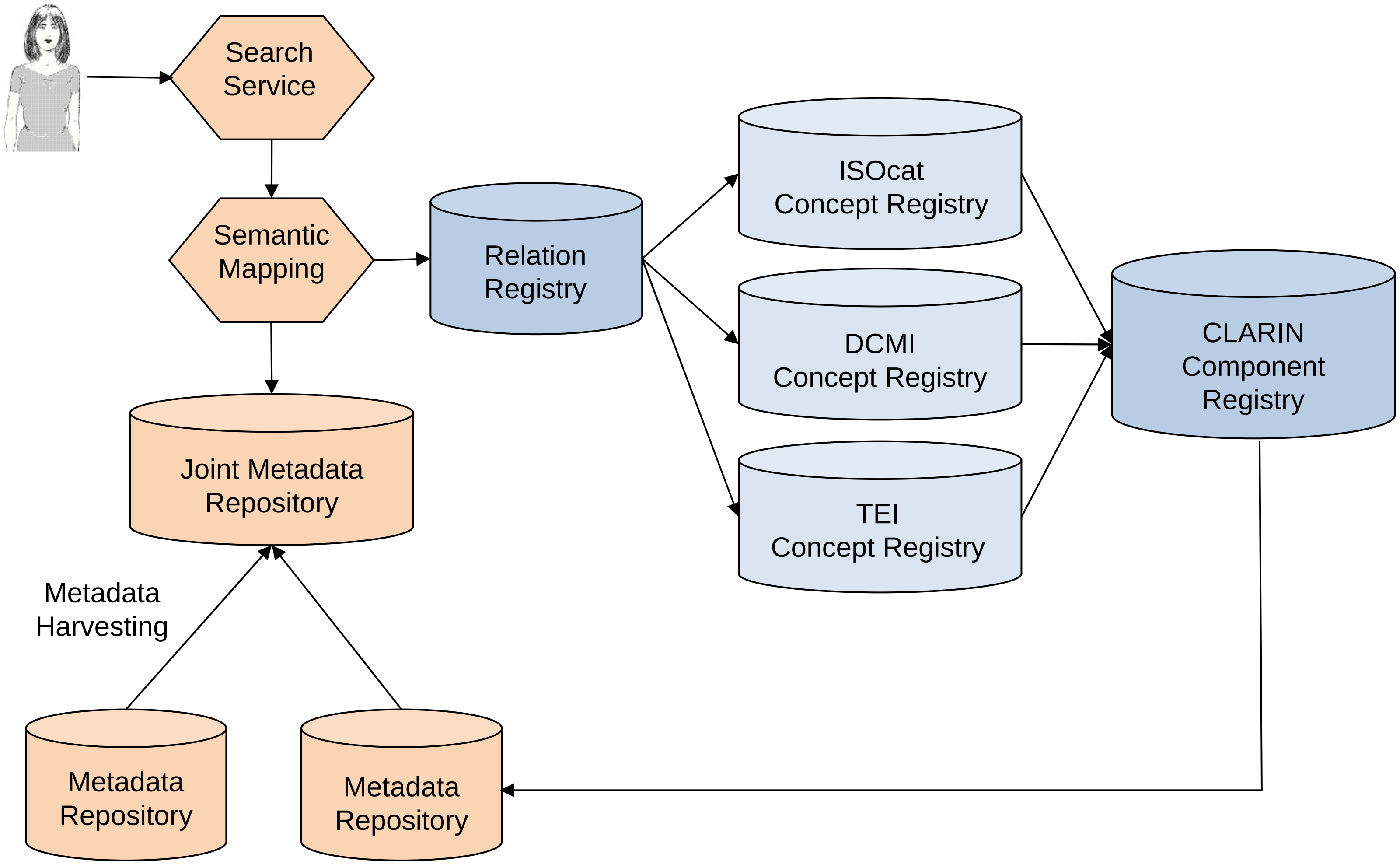
- CMDI is agreed after several meetings of various sorts (broad & small)
- current state and activities in two tracks - requirements doc is available
- track 1: element definitions
  - basic metadata categories have been determined for resources and tools/services
  - ISOcat (ISO 12620/ISO 11179) framework is stable to register all concepts
  - ws expert groups are working - elements are open for comments
- track 2: infrastructure
  - component specifications are available (zip file at the WP2 site)
  - working group formed to develop software framework
  - framework with registries, portals, harvesters, editors, search/browsers, GIS overlays, etc
  - WG is open for others to contribute - but need solid developers

- CMDI is CLARIN standard - exceptions can’t be accepted
- working on a Virtual Language Observatory

CMDI component framework



CMDI infrastructure



Virtual Collection building

- first “simple” step is integration:

  allow people to create a virtual collection by combining resources from different resource providers

- what are the ingredients?
  - joint metadata domain (working on that, harvesting via OAI and XML/HTTP)
  - single identity/single sign-on domain

    (working on this together with eduGain/TERENA

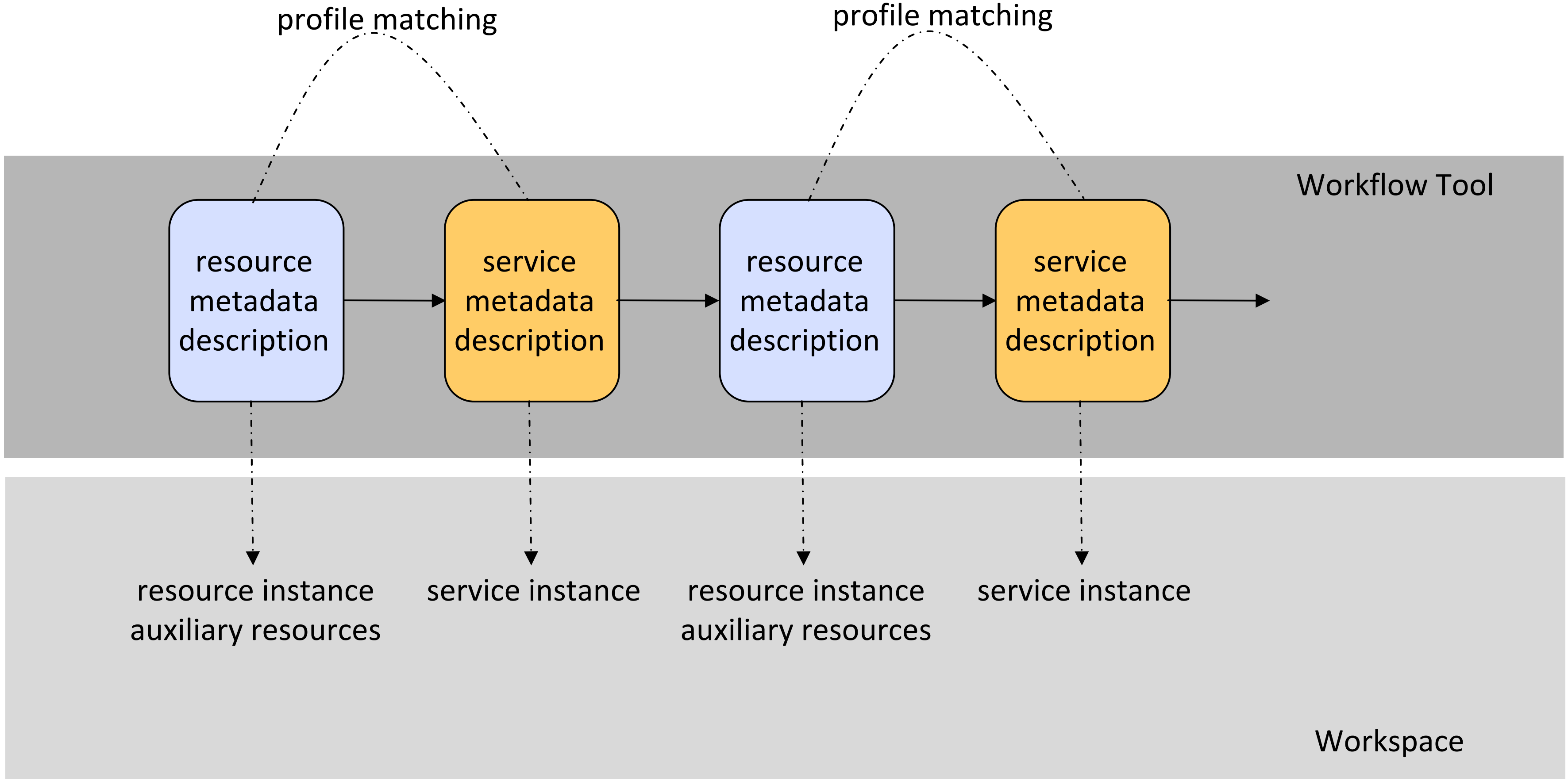
    probably now a first testbed with Dutch, German & Finnish institutions)
    - CLARIN centres will act as a “Service Provider Federation” , i.e. working on agreements
  - persistent identifier domain based on robust services

    MPG decided to support this at GWDG - should be open for research

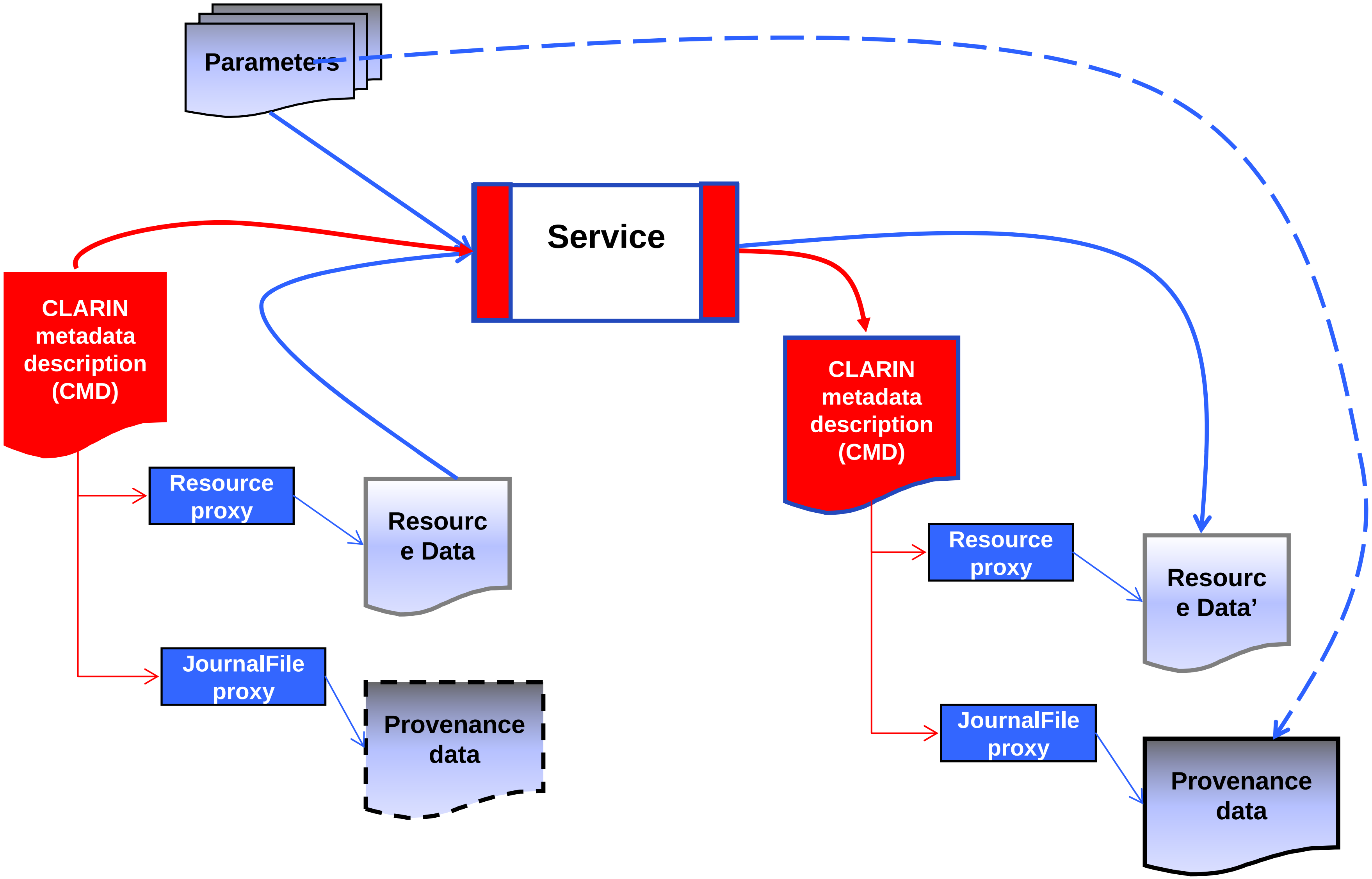
    basis is the Handle System and additional functions

Workflow building

- next step is to allow users to create workflows
- architecture is kind of clear - also MD profile matching principles


profile matching
profile matching
Workflow Tool
resource metadata description
service metadata description
resource metadata description
service metadata description
resource instance auxiliary resources
service instance
resource instance auxiliary resources
service instance
Workspace

MD in workflows



but interoperability ...

- most difficult problems - just a few comments

- three major aspects:
  - basic encoding (UNICODE, lin PCM, JPEG, MPEG, etc)
    - taken care of by large discipline crossing communities
    - still much dynamics in video encoding and archiving (->lossless MJPEG2000)
  - formatting - resource structuring (XML just the agreed language)
    - fairly regular for time series of all kinds
    - tricky for semi-structured data (lexica, complex annotations, text documents, etc)
    - working towards more generic formats - of course less specificity
    - most generic format is RDF assertions - but loss of any syntactic compactness

  - encoding of phenomena

but interoperability ...

- three major aspects:
    - basic encoding (UNICODE, lin PCM, JPEG, MPEG, etc)
    - formatting - resource structuring (XML just the agreed language)
    - encoding of phenomena
        - this is the result and/or preparation of research
        - very much theory and intention dependent
        - what does interoperability mean and where is it for????
        - domain ontologies will work where difference is just in terminology and where classification systems are stable
        - in our domain we just started with data category registry based on ISO 12620 as a reference (all based on ISO 11179)

          on purpose we left the relations out of any harmonization efforts

Cost aspects

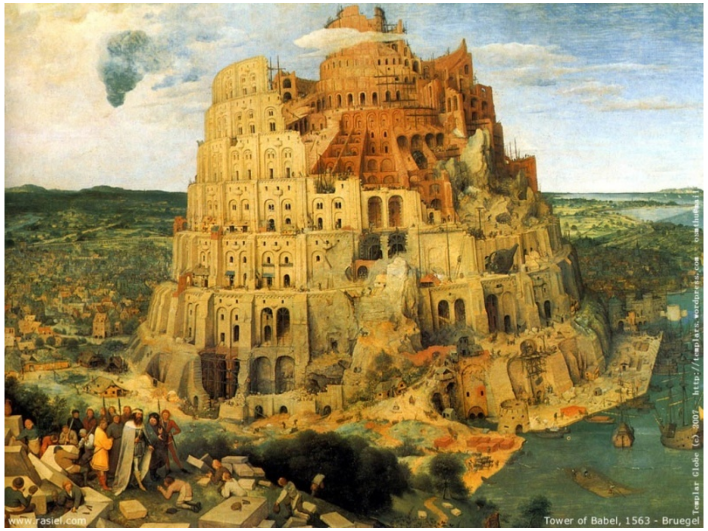
- Beagrie:
  - acquisition&ingest (43%), storage&preservation (23), access (35)
  - after 10 years metadata creation costs are factor 10 more expensive
- Dimper: disc capacity doubles every 13 months - data volume doubles every 15 months
- MPG: costs of current volume is 10% of costs after storage innovation cycle (10y)
- MPI: maintaining a complex language archive (50 TB, 600.000 objects)

  own repository (80 k€), 4 copies at CC (10 k€), system&archive manager (120 k€)

  archive & access software maintenance (180 k€)
  - economy of scale: more data could be managed
- do we want to give all our gold to Google or MS clouds?
  - which costs would be reduced - which not? what would it solve?
  - CCs are not very expensive

End

Falls nicht to end in Babylonish scenario nous avons still een beten time om mechanismes te improve.


www.rasief.com
Tower of Babel, 1563 - Bruegel
Templar Globe (c) 2007 http://templars.wordpress.com

Thanks for your attention!

NEERI 09
1/2. October in Helsinki
http://www.csc.fi/english/pages/neeri09