University of Minho

Open Repositories 2007
Plenary Session 2

# CRiB
## Preservation Services for Digital Repositories


Miguel Ferreira
mferreira@dsi.uminho.pt

Ana Alice Baptista
analice@dsi.uminho.pt

José Carlos Ramalho
jcr@dsi.uminho.pt


January 25th, 2007

University of Minho

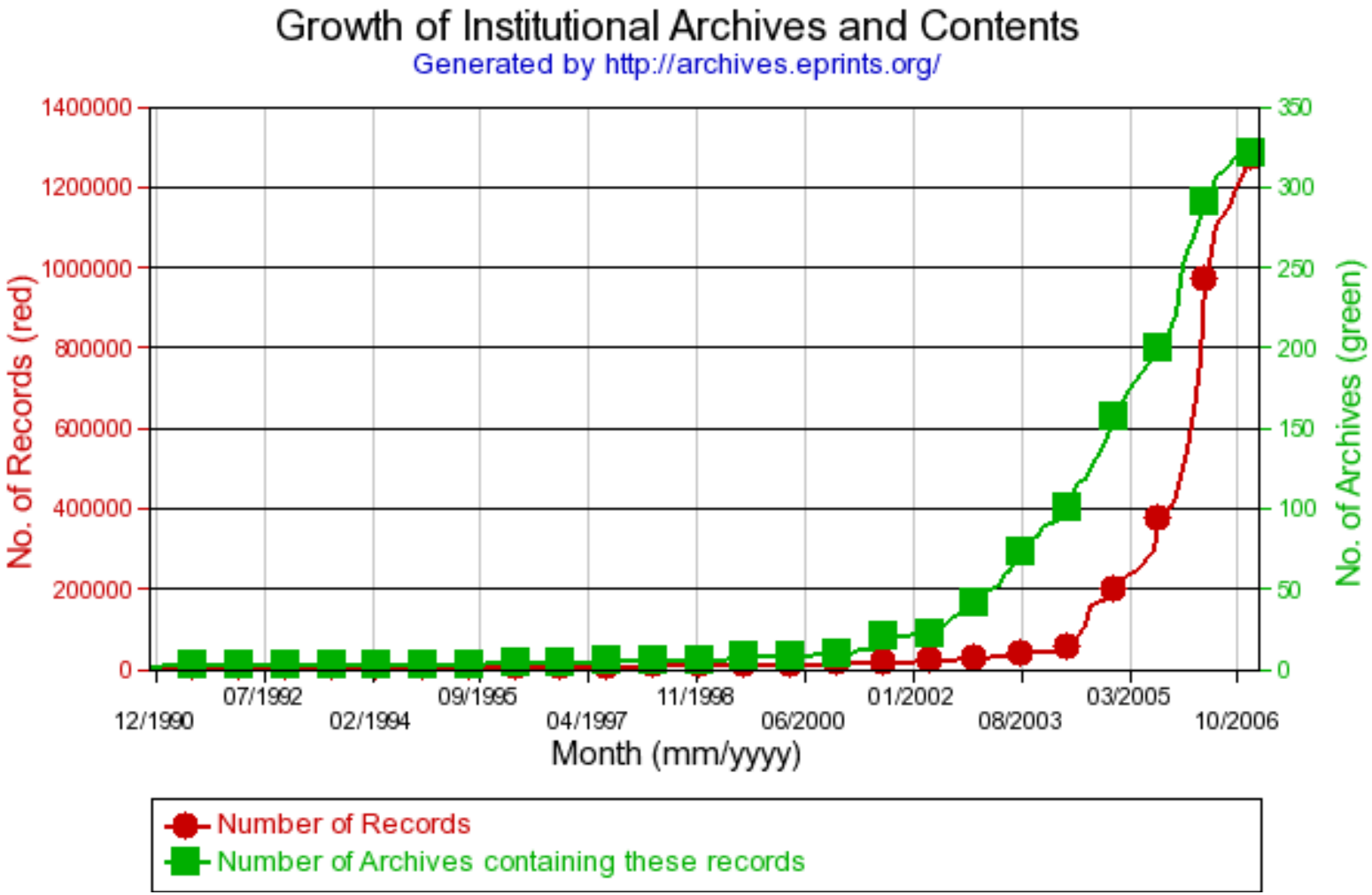
# Why are we using repositories?

- **Large production** of digital materials
  - Easy to create, great quality, very easy to disseminate
- **Affordable** technology
  - e.g. Eprints, DSpace, Fedora
- Take less storage **space** than analogue materials
- Some materials can only exist in digital form
  - e.g. Web site, 3D model, relational database, flash interactive animation
- Exponential growth of **adoption**

University of Minho

Open Repositories 2007
Plenary Session 2

# Adoption curve



# Repository limitations

- Excellent at **archiving** and **disseminating** materials
- **Poor at preserving** those materials in the long-run
  - Bit preservation
  - Normalization of formats during ingest
  - Store technical metadata
    - MD5 checksum, file format, ...
  - Supported formats list in DSpace
    - Supported, known, unsupported
  - Little preoccupation with authenticity

University of Minho

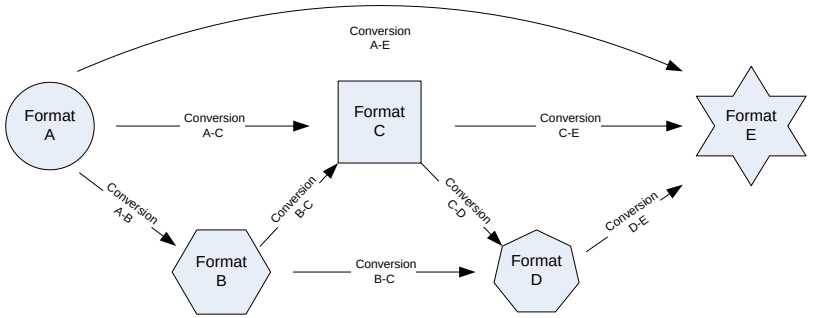
Open Repositories 2007
Plenary Session 2

# Digital preservation

- A definition
  - The set of **processes** and activities that ensure the **continued access to information** existing in **digital formats**
- Preservation strategies
  - Emulation
  - Encapsulation
  - **Migration***

# Distributed migration

- **Remote conversion services**
  - known APIs
  - descriptive metadata for localization and invocation (UDDI)
- Advantages
  - **Platform independency**
  - **Redundancy/multiple** migration paths
  - Compatible with other migration strategies
    - Normalization, migration on request
  - Generalized **cost reduction**
- Disadvantages
  - **Bandwidth** requirements
  - **Slow**
- Examples
  - PANIC
  - MyMorph (National Library of Medicine)
  - TOM (Typed Objects Model)



University of Minho

# What's the best preservation strategy?

- **Multiple** preservation **choices** available
  - Various formats, several converters for each pair of formats
- Lack of **universal acceptance** or objectivity
- Distinct preservation requirements
  - Satisfaction of the **designated community**
  - Characteristics of the **collection**
  - **Budget**
- Framework for evaluating preservation strategies [Rauch and Rauber]
  - Utility Analysis

University of Minho

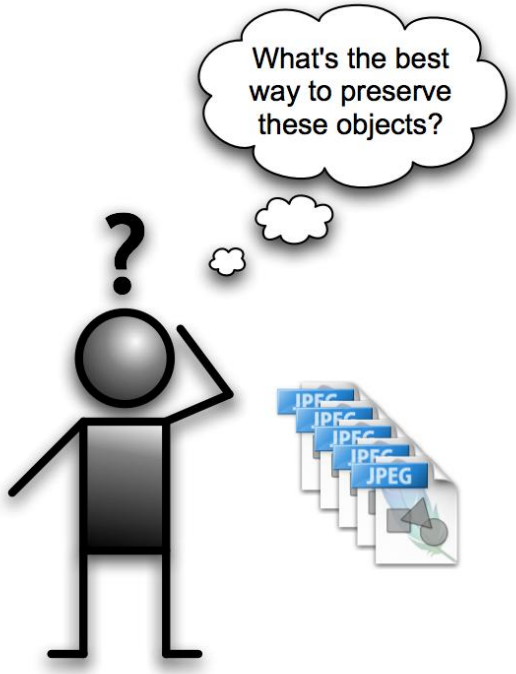
Open Repositories 2007
Plenary Session 2

# The CRiB platform

- Service Oriented Architecture (SOA)
- **Recommendation** service
  - Recommends an optimal migration strategy taking into account
  - Requirements of each client institution
  - Behavior/quality of each migration service
- **Migration** services
  - Service composition
- **Evaluates** the outcome of each migration
  - Performance, data loss, format characteristics
  - Produces an evaluation report (authenticity)

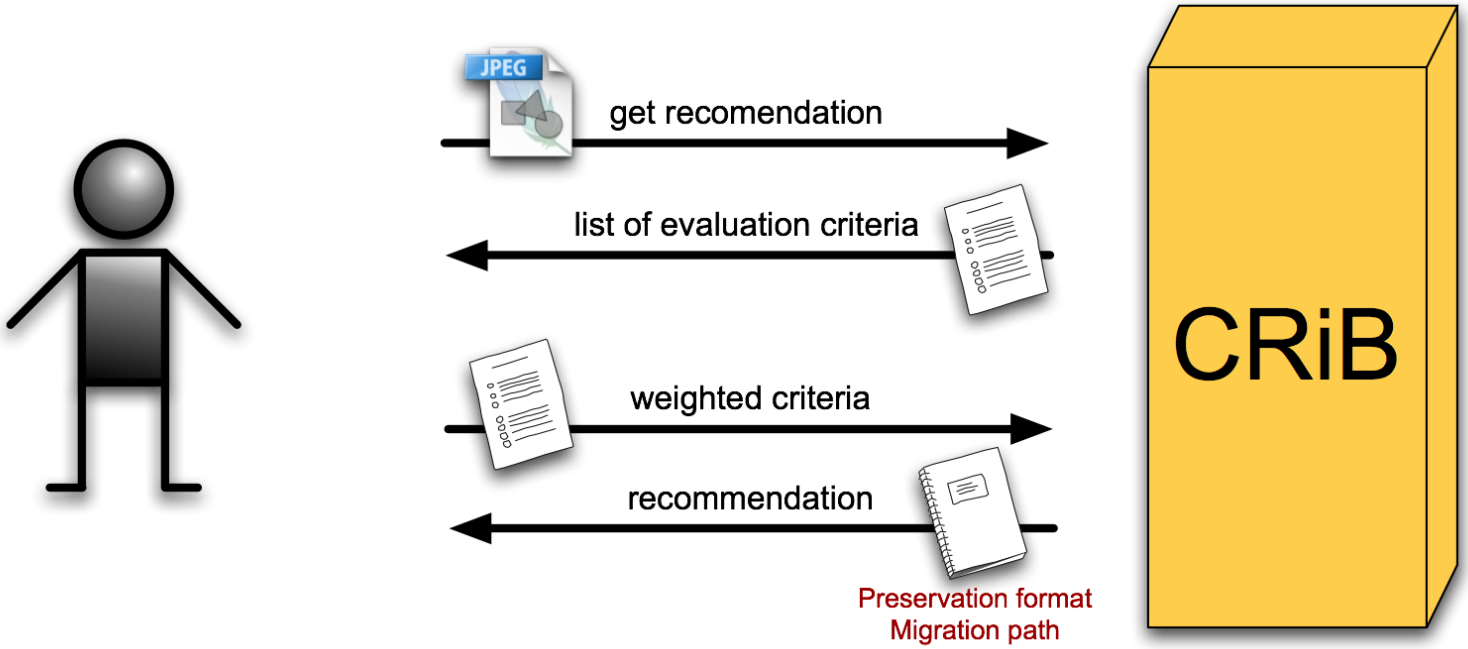
# Scenario

- A collection of digital objects of a certain format
  - e.g. JPEG files collected from a digital camera
  - e.g. A collection of text documents



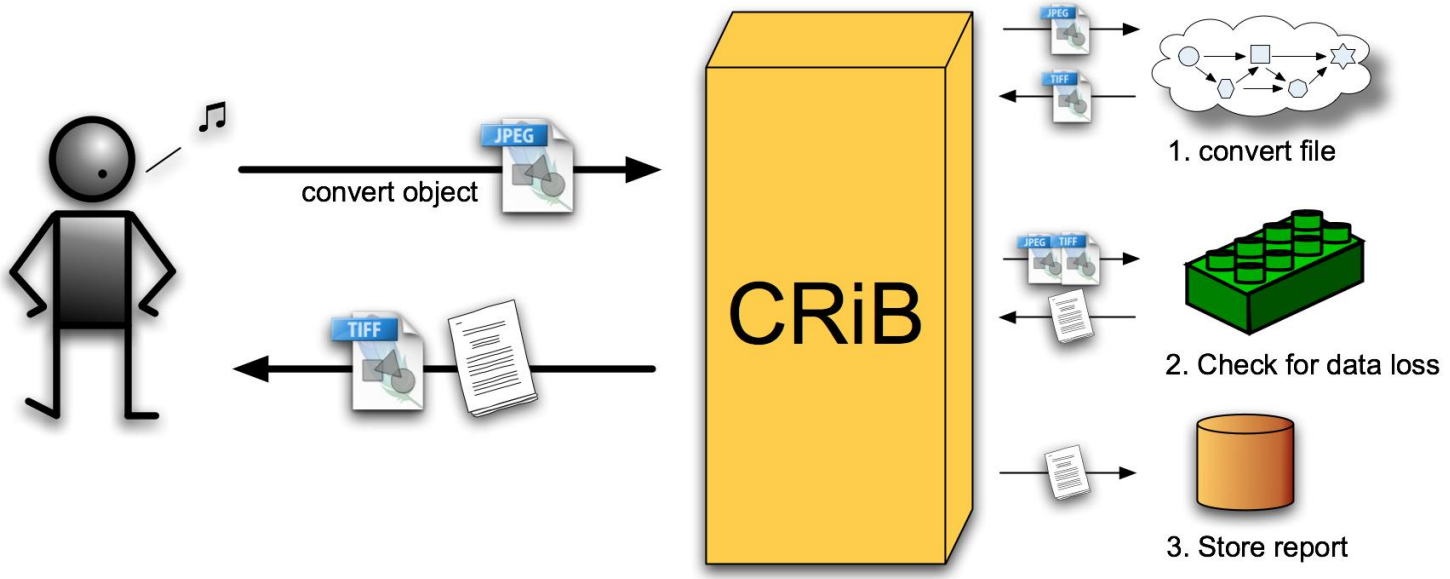
# Scenario

- Using the recommendation service
  - Preservation format (i.e. The target format)
  - Migration service (or combination of services)



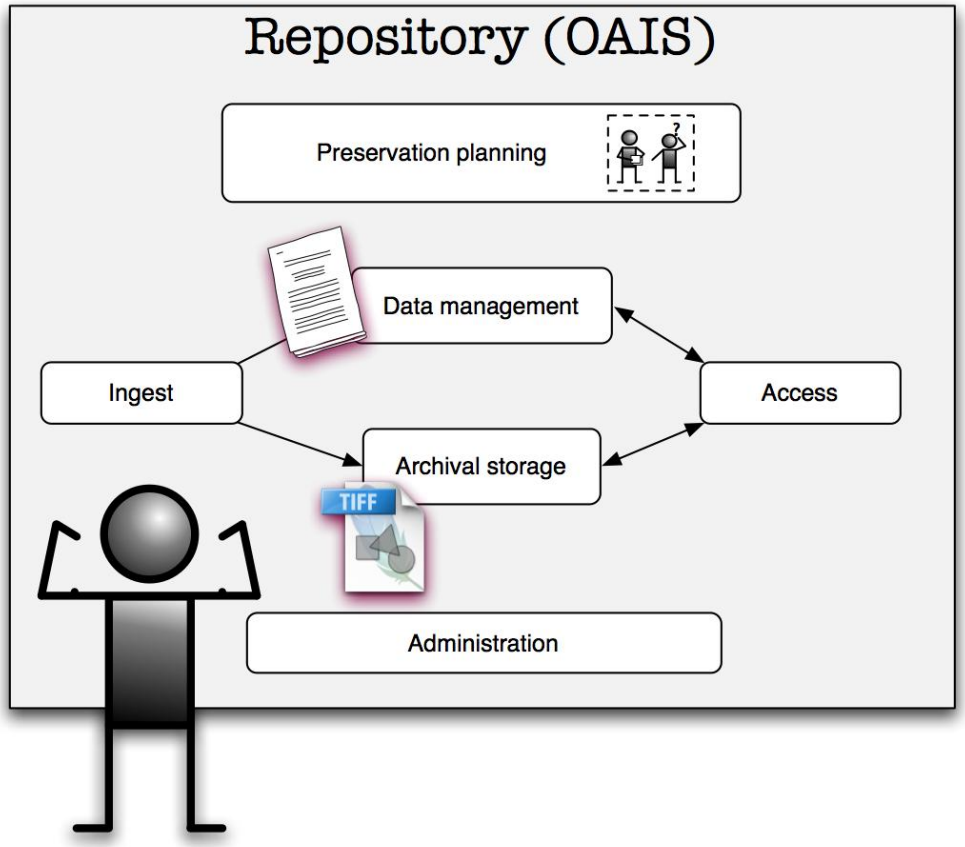
# Scenario

- Using the conversion services
  - Check for data loss and generate a migration report
  - Store the report
  - Return the converted file and the report back to the user

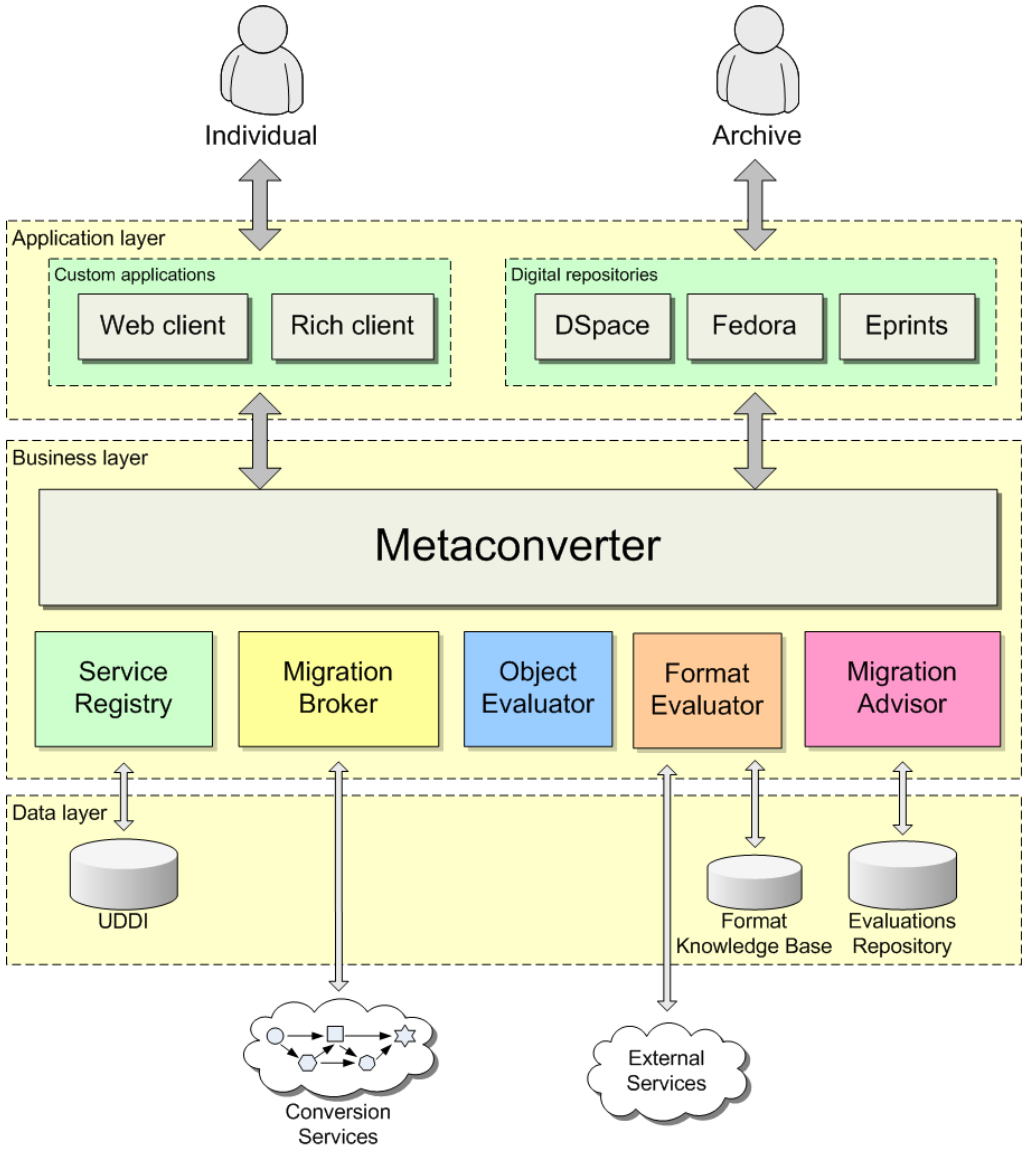


# Scenario

- Store the converted object
- Embed the metadata

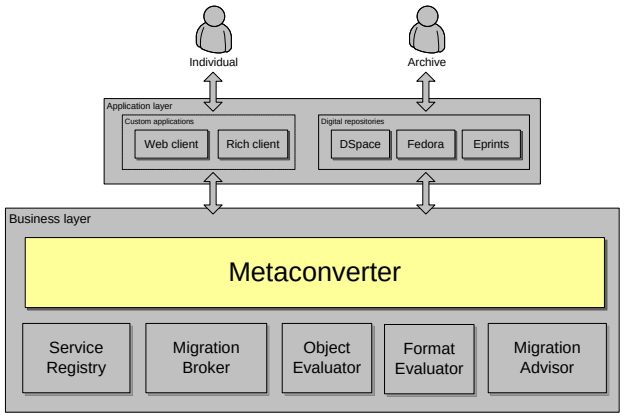


# Detailed architecture



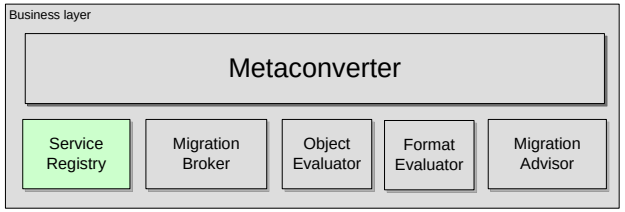
# Metaconverter

- Handles all **communication** between the **client** and the **CRiB** system
  - Its a web service
- Orchestrates the **communication** within the system and its **components**



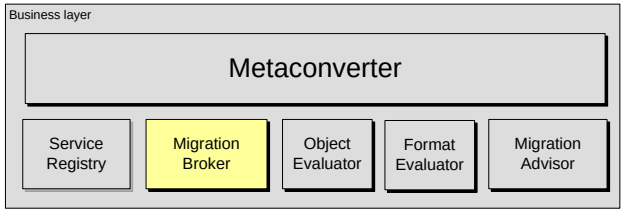
# Service Registry

- Manages information about conversion services
- Based on **UDDI**
  - Producer/developer information
    - Name, description, contact
  - Service information
    - Name, description, source/target formats, cost of invocation, ...
  - Binding information
    - How the service can be invoked
  - **Source/target** information
    - Controlled vocabulary based on PRONOM file format descriptors



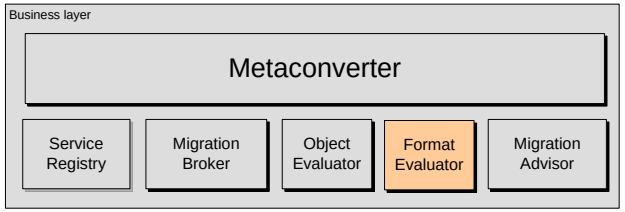
# Migration Broker

- **Carries out** format **conversions**
  - Invokes all the necessary conversion services
- Measures the **performance** of the conversion process
  - Availability
  - Stability
  - Throughput
  - Scalability
  - Cost
  - Size ratio
  - File count ratio



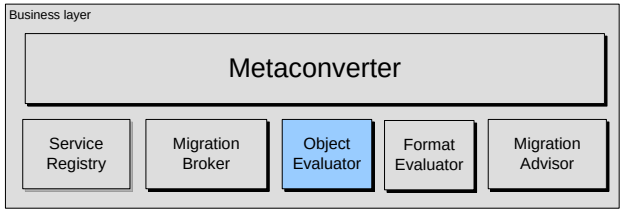
# Format Evaluator

- Provides useful information about the **status** of involved **formats**
  - Market share
  - Support level
  - Lossy compression only
  - Embedded metadata
  - Royalty-free
  - Backward compatibility
- Format Knowledge base
  - Database of facts about each format
  - PRONOM Registry*
  - Google trends*



# Object Evaluator

- Determines the amount of **data loss** involved in the migration
- Detects the **similarity** between the significant properties of digital objects
  - **Depends** on the **class** of **objects**
  - Different significant properties for bitmap images, text document, relational databases, etc.
- Produces **evaluation reports** in PREMIS format (eventOutcomeDetail)
  - **Datetime** of intervention
  - Description of involved **agents**
  - **Type** of event (i.e. Migration)
  - **Outcome** of the intervention



University of Minho

# Significant properties [still images]

**Table 3** Taxonomy of significant properties for still images.

| Criterion name | Description |
|---|---|
| appearance::resolution::width | The width of the digital image measured in pixels. |
| appearance::resolution::height | The height of the digital image measured in pixels. |
| appearance::colour::model | Abstract mathematical model describing the way colors can be represented as tuples of numbers, typically as three or four values or color components (e.g. RGB, sRGB, HSL, HSV, YUV, CMYK) [57]. |
| appearance::colour::depth | The number of bits used to represent the color of a single pixel in a bitmapped image or video frame buffer [56]. |
| content::completeness::pixel_correctness | How well pixels are respectful to the original image. In cases of multiple page images, the comparison is performed by page and an overall similarity value is calculated by averaging the whole page set results. |
| content::completeness::page_count | The number of pages that constitute the image. |
| context::metadata | Some image formats embed metadata. This criterion intends to measure how much of that metadata has been preserved. |
| structure::compression::method | Image compression can be lossy or lossless. Examples of lossless image compression methods are: run-length encoding, entropy coding and adaptive dictionary algorithms such as LZW. Examples of methods for lossy compression are: reducing the color space to the most common colors in the image, chroma subsampling, transform coding and Fractal compression [58]. |
| structure::compression::level | The level compression used in the object. The value zero is used when lossless compression methods are in place. |

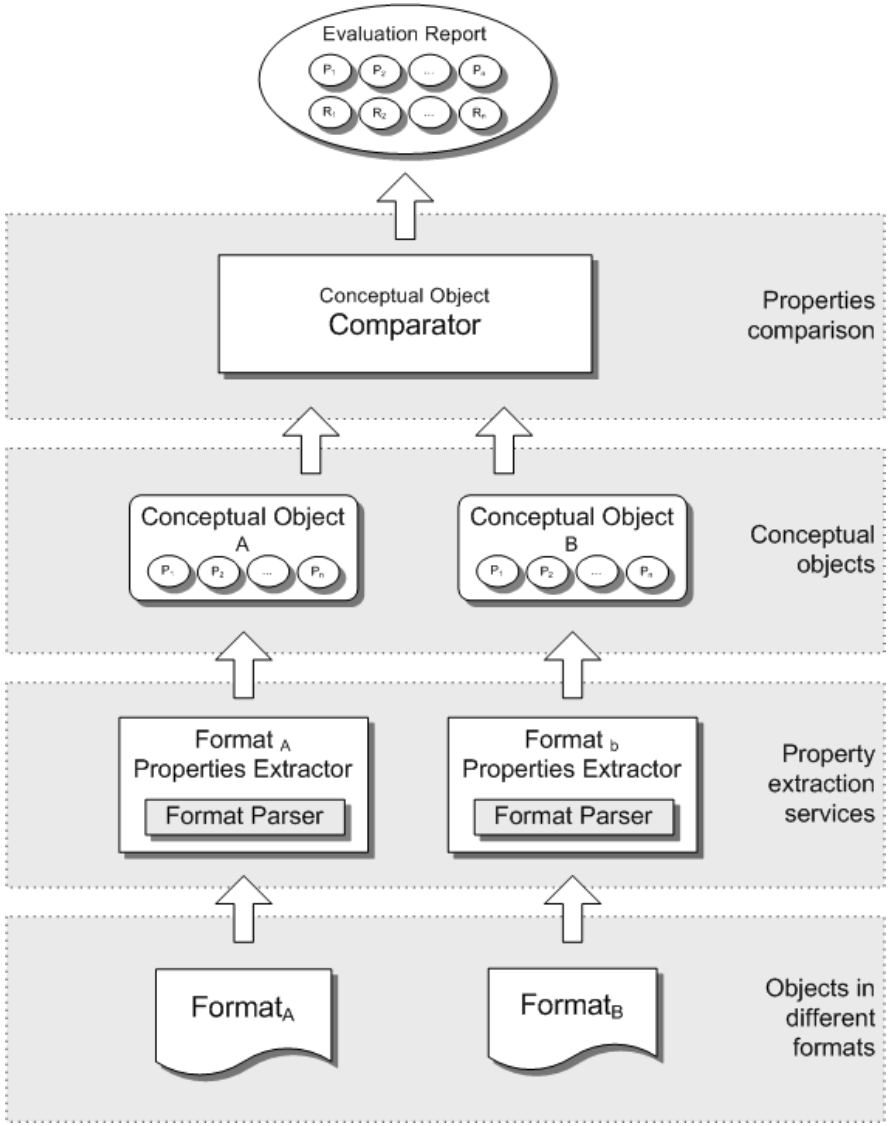
University of Minho

# Significant properties [text documents]

**Table 4** Taxonomy of significant properties for text documents.

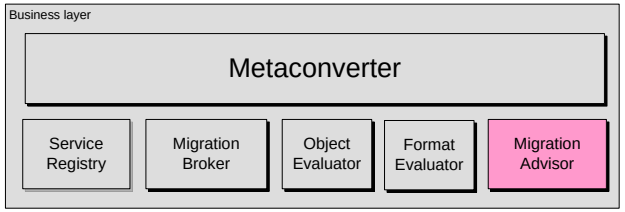
| Criterion name | Description |
|---|---|
| appearance::page::width | The width of the document measured in millimeters in relation to the original document. |
| appearance::page::height | The height of the document measured in millimeters. |
| appearance::page::layout | How similar the layout of the text-document is in relation to the original. |
| appearance::page::margins::left | The size of the left margin of the document. |
| appearance::page::margins::right | The size of the right margin of the document. |
| appearance::page::margins::top | The size of the top margin of the document. |
| appearance::page::margins::bottom | The size of the bottom margin of the document. |
| appearance::page::style::background color | The predominant background color of the document. |
| appearance::page::style::font faces | Collection of fonts used throughout the document. |
| content::completeness::character correctness | How well text characters are respectful to the original document. In cases of multiple page documents, the comparison should be performed by page and an overall similarity value calculated by averaging the whole page set results. |
| content::completeness::page count | The number of pages that constitute the text document. |
| content::completeness::image count | The number of images embedded in the text document. |
| context::metadata | Some text document formats carry embedded metadata. This criterion is expected to determine how much of that metadata has been preserved. |

# Object evaluator under the hood

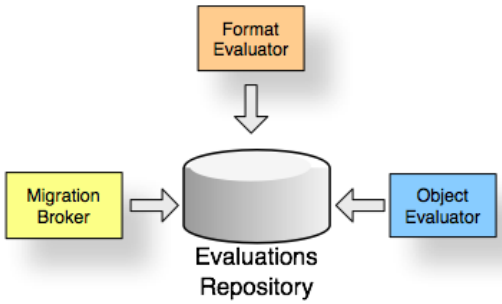


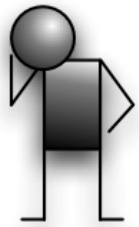
# Migration Advisor

- Generates **recommendations** of optimal migration choices
- Uses information provided by the **client** to determine the **best available option**
  - Clients weight each of the evaluation criteria according to their personal requirements
- Confronts those **requirements** with the **accumulated knowledge** about the behavior of each conversion service
  - Performance
  - Data loss
  - Format status



# Recommendation engine



client

Migration Advisor

| | Evaluation criteria | WS1 | WS2 |
|---|---|---|---|
| Process | Availability | 95% | 90% |
| Process | Stability | 80% | 90% |
| Process | Scalability | 1 | 1.5 |
| Process | Throughput | 1 Kb/s | 2 Kb/s |
| Process | Cost | 0.01 € | 0.05 € |
| Object | Number of chars | 70% | 95% |
| Object | Ordering of chars | 60% | 85% |
| Object | Page layout | 80% | 20% |
| Format | Market share | 30% | 70% |
| Format | Support level | 5 | 4 |
| Format | Open standard | yes | no |

| Weights |
|---|
| 40% |
| 30% |
| 0% |
| 20% |
| 10% |
| 40% |
| 40% |
| 20% |
| 30% |
| 30% |
| 40% |

| WS1 | WS2 |
|---|---|
| 5 | 5 |
| 5 | 5 |
| 1 | 5 |
| 1 | 5 |
| 5 | 1 |
| 4 | 5 |
| 4 | 5 |
| 5 | 2 |
| 2 | 4 |
| 5 | 4 |
| 5 | 1 |

x

| Weights |
|---|
| 0.4 |
| 0.3 |
| 0 |
| 0.2 |
| 0.1 |
| 0.4 |
| 0.4 |
| 0.2 |
| 0.3 |
| 0.3 |
| 0.4 |

=

| WS1 | WS2 |
|---|---|
| 2 | 2 |
| 1.5 | 1.5 |
| 0 | 0 |
| 0.2 | 1 |
| 0.5 | 0.1 |
| 1.6 | 2 |
| 1.6 | 2 |
| 1 | 0.4 |
| 0.6 | 1.2 |
| 1.5 | 1.2 |
| 2 | 0.4 |
| 12.5 | 11.8 |

(1) Average evaluations per service

(2) User assigned weights of importance

(3) Normalised evaluations [1-5]

(4) Ranking process

(5) Overall rating

University of Minho

Open Repositories 2007
Plenary Session 2

# Round-up

- Platform for **executing**, **evaluating** and **recommending** migration-based preservation interventions
- Produces **PREMIS** metadata reports
  - Document the intervention (**eventOutcomeDetail**)
  - Important for authenticity
- Reduction of preservation costs
  - Broad range of converters available
  - Recommendation service enables automatic preservation planning
    - still needs an obsolescence notifier

University of Minho

Open Repositories 2007
Plenary Session 2

# Round-up

- **Extensible**
  - Possibility of adding new conversion services and evaluators
- **Platform independent**
- Objective way of **benchmarking** of converters
- Enables the community to **cooperate** by:
  - Publishing new conversion services
  - Developing similarity algorithms for such properties
    - Necessary for the Object Evaluator

University of Minho

# Current status & future work

- All components are developed and ready for testing
  - Finishing the **integration of evaluators**
  - **Demo** at the Project webpage
    - Migration Workbench
- Evaluation
  - **Cross validation** on the Migration Advisor
  - Raster **images**
    - PNG, BMP, TIFF, GIF, JP2, JPEG
  - **Text** documents
    - Word, OpenDocument (ODT), PDF, RTF
- Future work
  - Handle **more formats** and object classes
  - Enrich **evaluation taxonomies**
    - INSPECT Project?

University of Minho

Open Repositories 2007
Plenary Session 2

# Questions?

CRiB

Conversion and Recommendation of Digital Object Formats

More information at http://crib.dsi.uminho.pt


Miguel Ferreira
mferreira@dsi.uminho.pt

Ana Alice Baptista
analice@dsi.uminho.pt

José Carlos Ramalho
jcr@dsi.uminho.pt