GRANT AGREEMENT: 601138 | SCHEME FP7 ICT 2011.4.3
Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics [Digital Preservation]

Semi-automated metadata extraction in the long term

DPC Workshop, Belfast, Dec 2015

“This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no601138”.

Structure of presentation

- Introduction to Pericles
- Layers of Metadata
- Sources of Metadata
- Time, space and data
- Semi-automated metadata as mitigating factor

Introduction to Pericles

Introduction to PERICLES

- Four-year Integrated Project (2013-2017) funded by the European Union under its Seventh Framework Programme
- **P**romoting and **E**nhancing **R**euse of **I**nformation throughout the **C**ontent **L**ifecycle taking account of **E**volving **S**emantics
- Two domains:
  - Digital artworks, such as interactive software-based installations, and other digital media from Tate's collections
  - Material from Tate's archives
  - Experimental scientific data originating from the European Space Agency and International Space Station.

Model-driven approach

- Essentially all archives are based around some conceptual model of the material held
- PERICLES applies formal models to describe
  - Objects
  - Entities associated with objects
  - Broader community
- These models support processes such as **appraisal** and **QA**, and consequentially functionality such as maintenance and actions taken for sustainability
- A broad variety of models are under consideration: semantic (ontological) models to formally describe objects; social network graphs to describe community; statistical models to describe technology obsolescence...

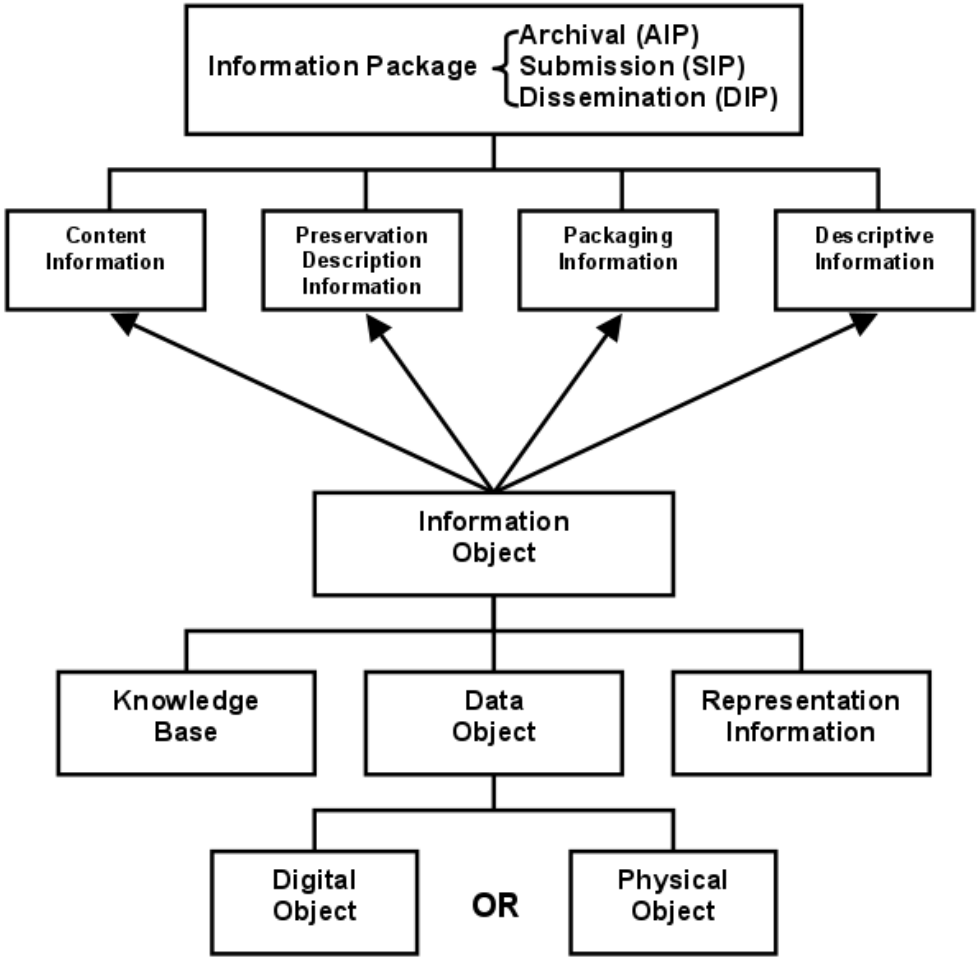
Layers of Metadata

Open Archival Information System

- OAIS reference model
  - *“conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term“*

    -Lavoie, B. (2000). Meeting the challenges of digital preservation: The OAIS reference model
- OAIS-compliance
  - adherence to ISO 14721:2003 or (now) ISO 14721:2012
  - Specifies conceptual framework, functional model, information model

OAIS information model



Descriptive metadata

- Supporting humans and machines
- Goal: interpreting data object
- Not always possible to automatically interpret data objects on any level (some are fully opaque)
- Consider:
  - 'Unstructured' natural-language texts, such as letters, books, articles
  - Images of artworks
  - Images of letters
  - Recordings of audiovisual presentations
  - Complex data files

Sources of descriptive metadata

Automated metadata extraction

- Popular view on indexing metadata:
  - “the more, the merrier”
- Risks of low-quality metadata:
  - Low accuracy on search and browse tasks; occasionally embarrassing misinterpretations
- Benefits:
  - Additional metadata can improve search indexing

How good is automated metadata extraction?

- Varies significantly depending on the precise task and source material
- Automated metadata extraction tends to apply probabilistic (machine learning) or heuristic approaches
- Machine-eye view:
  - describe what is present
  - Infer what is not based on:
    - Knowledge base
    - Comparison with other items
    - Learning from training examples ('supervised learning')

Crowdsourcing metadata

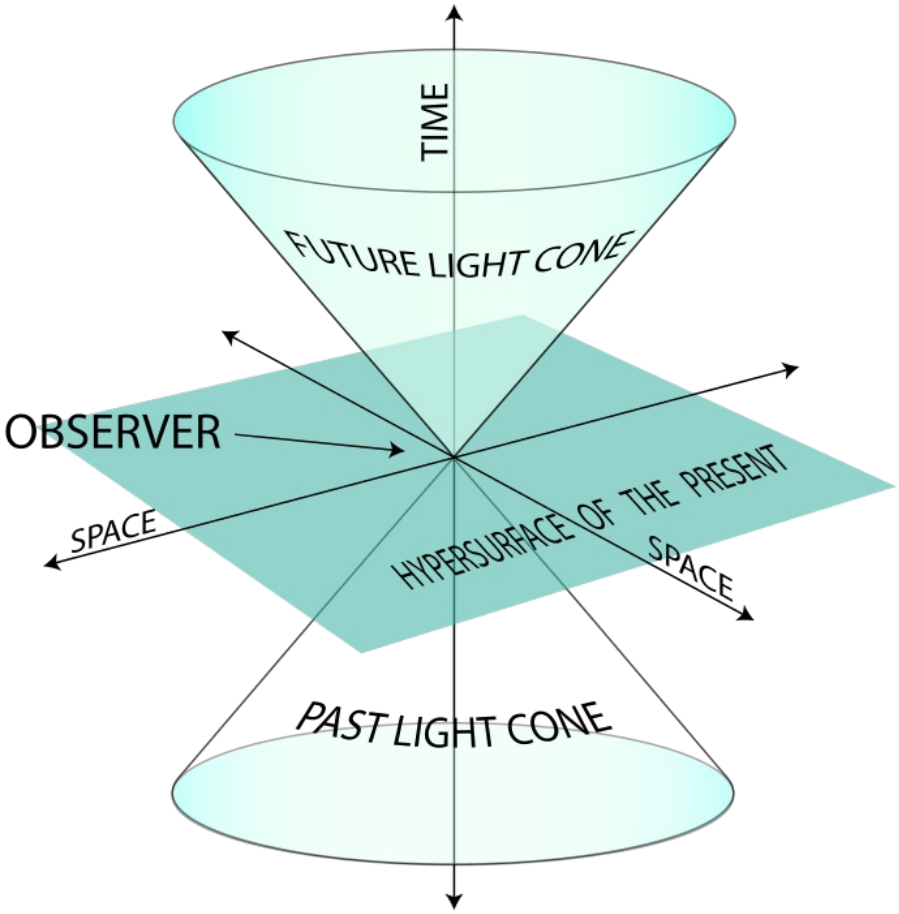
- The 'phone a friend' approach to metadata generation
  - Make material available to public
  - Encourage them to annotate (example: social tagging)
  - Examine the result
- The likely result:
  - Some material extensively annotated; some descriptive annotations; some formally structured; some personal ('cryptic')
  - Some/most material receives no notice and is not annotated at all
- Mitigation: engineer more consistent coverage through, for example, gamification (see Galaxy Zoo)
- Identify incentives that encourage public to contribute

Capturing 'live' metadata

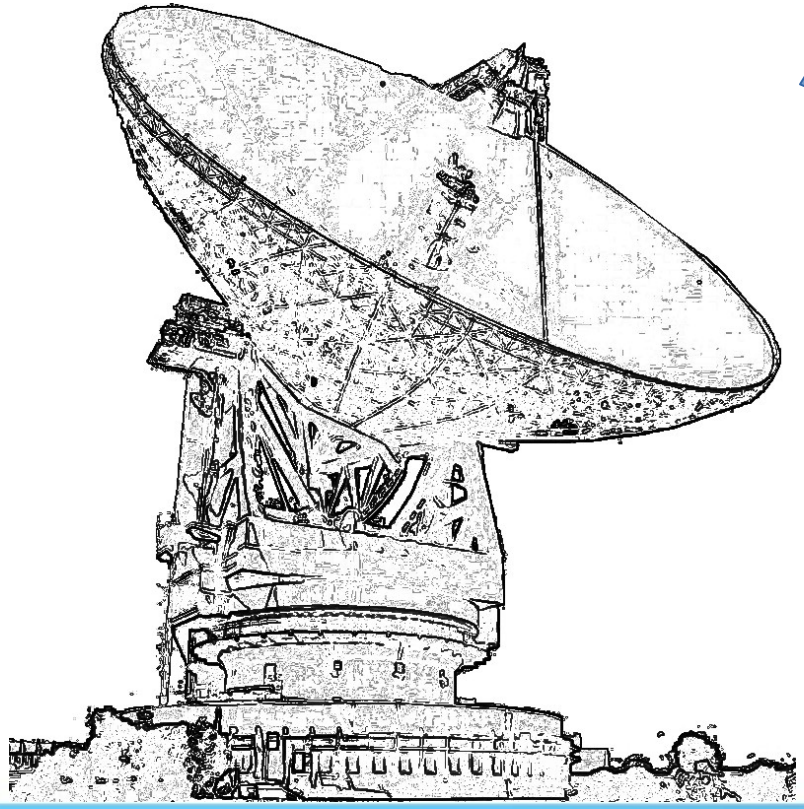
- If the environment is accessible at the time of creation:
  - Technical 'live' metadata may be captured
  - Within PERICLES, this is referred to as 'significant environment information'
  - Example: steps in creation, time of creation, contextual relevance of other files…
- Another sort of 'live' metadata emerges from observation of behaviour of those engaging with the data
  - Interaction with search/browse interfaces (cf. information scent)
  - Satisfaction with results
  - Patterns of sharing and reuse (information diffusion on social networks, for example)

Time, space and data

Theoretical reach of information



Theoretical reach of information



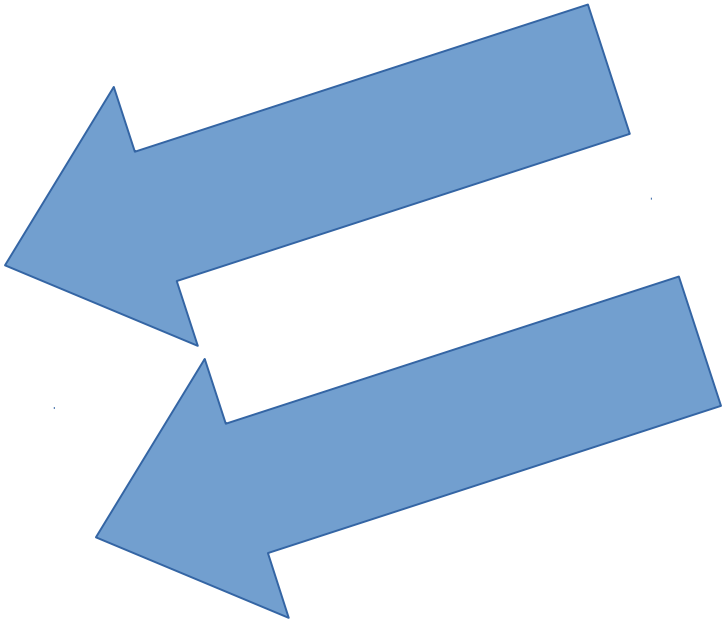
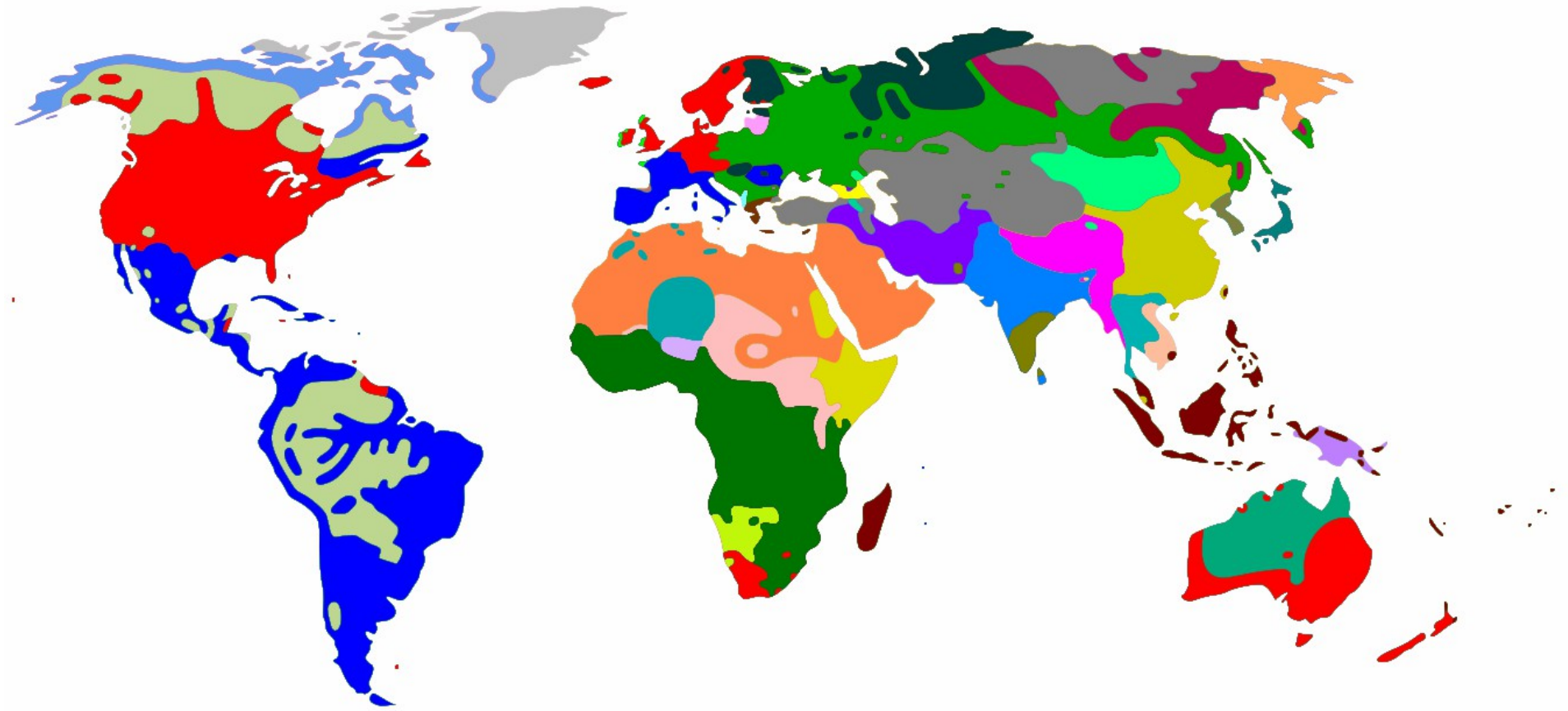


Image source: S Korotkiy

Practical reach of information

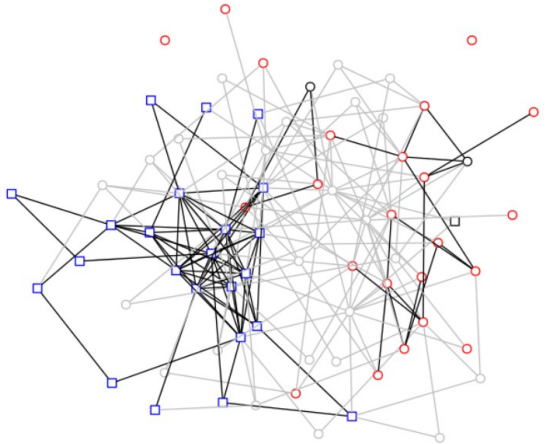
- Receiving the signal is only the start
- Can we decode the signal?
  - Technical decoding
  - Practical comprehension
- Confounding factors in decoding metadata:
  - Language
  - Dialect
  - Prerequisite knowledge

Language: space travel



Language change: Time travel

- Language may be viewed as a complex adaptive system (Beckner et al, 2007)
  - Made up of many tiny parts - people talking, writing, gesturing
  - Adaptive, because we change our behaviour based on past interactions
  - Many factors influence its development: biology of perception; social structure; experience
- Probabilistic processes underlie language change: collective experience and eventual consensus



Example: Photogram (Getty Art & Architecture Thesaurus)

photograms (<photographs by processing or presentation technique>, <photographs by technique>, ... Visual and Verbal Communication (hierarchy name))

Note: Photographs produced without a camera, usually by placing an object directly on sensitized paper and exposing it to light. Includes Talbot's first photographs, which he called photogenic drawings. In the early 20th century the term was sometimes used to mean all photographs.

Terms:

photograms (**preferred**,C,U,LC,English-P,D,U,PN)
photogram (C,U,English,AD,U,SN)
cameraless images (C,U,English,UF,U,N)
cameraless photographs (C,U,English,UF,U,N)
lensless photographs (C,U,English,UF,U,N)
photographs, cameraless (C,U,English,UF,U,N)
photographs, lensless (C,U,English,UF,U,N)
physiotypes (C,U,English,UF,U,N)
rayograms (C,U,English,UF,U,N)
Rayographs (C,U,English,UF,U,N)
Schadograms (C,U,English,UF,U,N)
Schadographs (C,U,English,UF,U,N)
shadowgraphs (photograms) (C,U,English,UF,U,N)
shadow pictures (C,U,English,UF,U,N)
黑影照片 (C,U,Chinese (traditional)-P,D,U,U)
光影圖像 (C,U,Chinese (traditional),UF,U,U)
實物投影 (C,U,Chinese (traditional),UF,U,U)
hēi yǐng zhào piàn (C,U,Chinese (transliterated Hanyu Pinyin)-P,UF,U,U)
hei ying zhao pian (C,U,Chinese (transliterated Pinyin without tones)-P,UF,U,U)
hei ying chao p'ien (C,U,Chinese (transliterated Wade-Giles)-P,UF,U,U)
fotogrammen (C,U,Dutch-P,D,U,U)
fotogram (C,U,Dutch,AD,U,U)
Fotogramme (C,U,German,D,PN)
Fotogramm (C,U,German-P,AD,SN)
Photogramm (C,U,German,UF,SN)
Photogramme (C,U,German,UF,SN)
fotogramas (otografias por técnica) (C,U,Spanish-P,D,U,PN)
fotograma (fotografía por técnica) (C,U,Spanish,AD,U,SN)

The challenges of decreasing accessibility

Understanding unfamiliar material

- Unfamiliar data
  - Technical encoding – well-understood problems
  - Challenges of internationalisation
- Unfamiliar texts
  - Conventions and best practices change over time
  - Coherence degrades long before it fails entirely (slower to read: takes more effort: machines trained on modern texts are likely to encounter issues with texts outside that timeframe)
- Challenges of unfamiliar artefacts
  - There are many more questions that may be asked about an object: for example, in the case of artworks, “artist's intent” may be significant
  - Once lost, these are very difficult to infer

Term recognition vs generation

- Understanding unfamiliar material, though hard, is *easier* than finding it
- Separate processes:
  - Recognising a term
  - Identifying (generating) a term
- Recognition is faster and more reliable
- Why:
  - Recognising a term: connecting term to concept
  - Generating terms: search around a concept looking through large pool of candidate terms for the one that might work best here
  - Think yourself into the curator's shoes: what terms might they have used for the concept that interests you, and why?

Semi-automated metadata as mitigating factor

Relating concept, feature, agent and term

- Peirce: semiotic triad, relating symbol, object and interpreter
  - Software agents: machine-level **features** (machine perception) – words found in documents, colours, shapes or patterns found in images…
  - Human agents: perception; comprehension; application of relevant knowledge; interpretation into a set of **concepts**; encoding observations into **terms**
- Observing the behaviour of human agents throughout the lifecycle of the digital object allows us to study change in **manual interpretation** and **encoding**
- This permits us to characterise these patterns of change
- It also permits software agents to be brought into line with changing norms

Conclusion

Conclusion

- In order to mitigate challenges of social, cultural, technology and semantic change, PERICLES combines
  - model-led approaches to data management
  - data-led approaches to modelling and characterising the changing environment and context(s) of reuse
- Approach acknowledges dynamical nature of system in which reuse occurs
- Downside: such an approach requires ongoing availability of material (ethically) gleaned from observational data
  - Consequentially, a closed archive or an archive that excites little interest remains difficult to sustain, unless data is sourced elsewhere
- In conclusion, therefore, data-led approaches gain from joint infrastructure and open data