ARKive-ERA Project Lessons and Thoughts

Semantic Web for Scientific and Cultural Organisations
Convitto della Calza
17th June 2003

Paul Shabajee (ILRT, University of Bristol)

Contents…

- Context – Digitisation Projects Worldwide
- Goals – Repurposing in Educational Settings
- ARKive-ERA Project & Requirements
- Some Key Problems Identified
- Why Semantic Web and Other Tools?
- Conclusion

Digital Collections of Multimedia

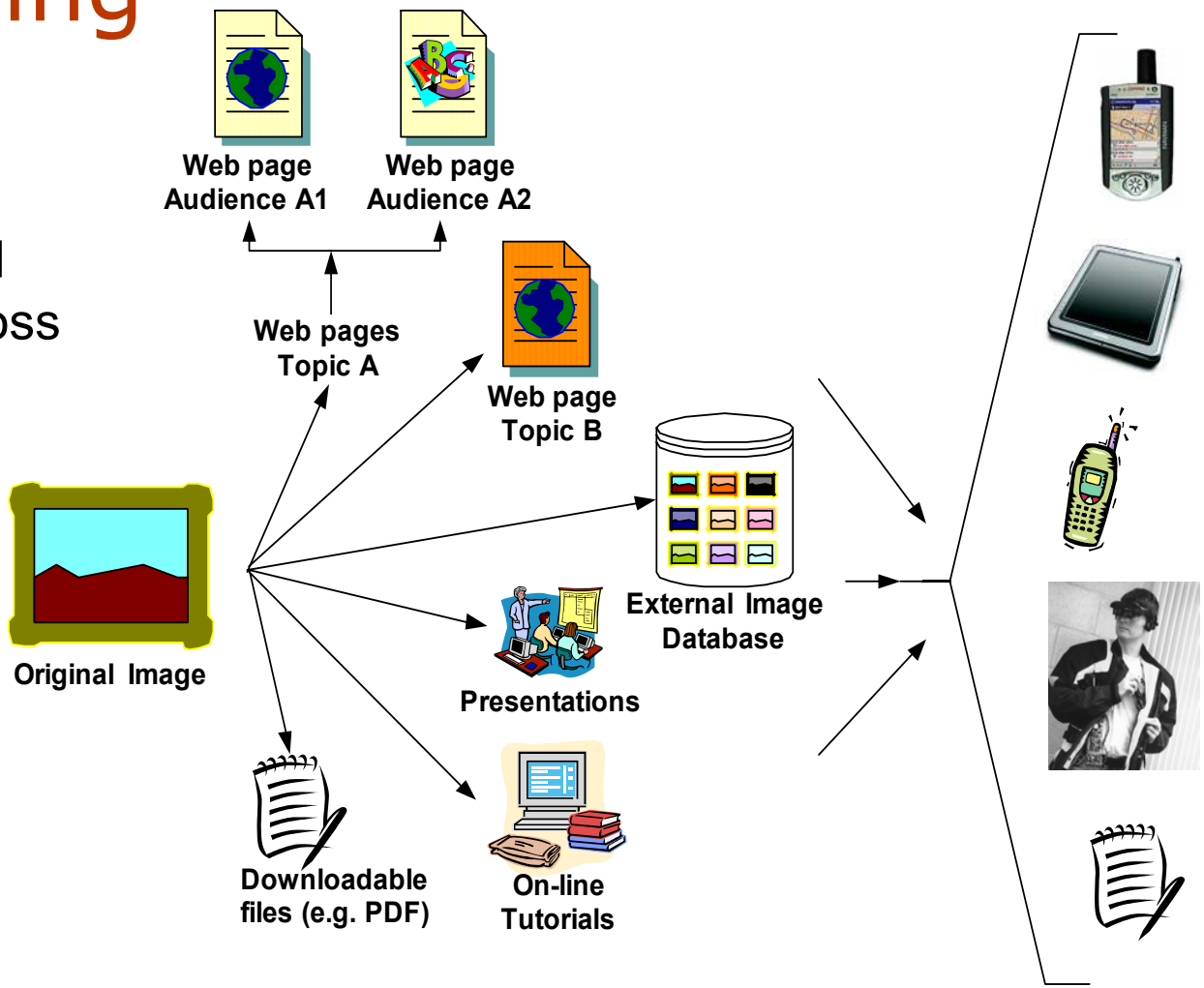
- Millions of £, € & $ invested in development of Very Large Digital Collections of…
  - Priceless Historic and Cultural Objects, Images and Films, 3D Environments, Places and Things Otherwise Inaccessible
  - Future… immersive 3D environments, 'wearable computing', new forms of collaboration and interaction…
  - Imagine: If these can be used to support learning and teaching in any relevant subject, phase of education, level of education, utilised to help provide personalised learning resources for individuals….
  - Want maximum 'educational' value from of digital collections

Digital Collections of Multimedia

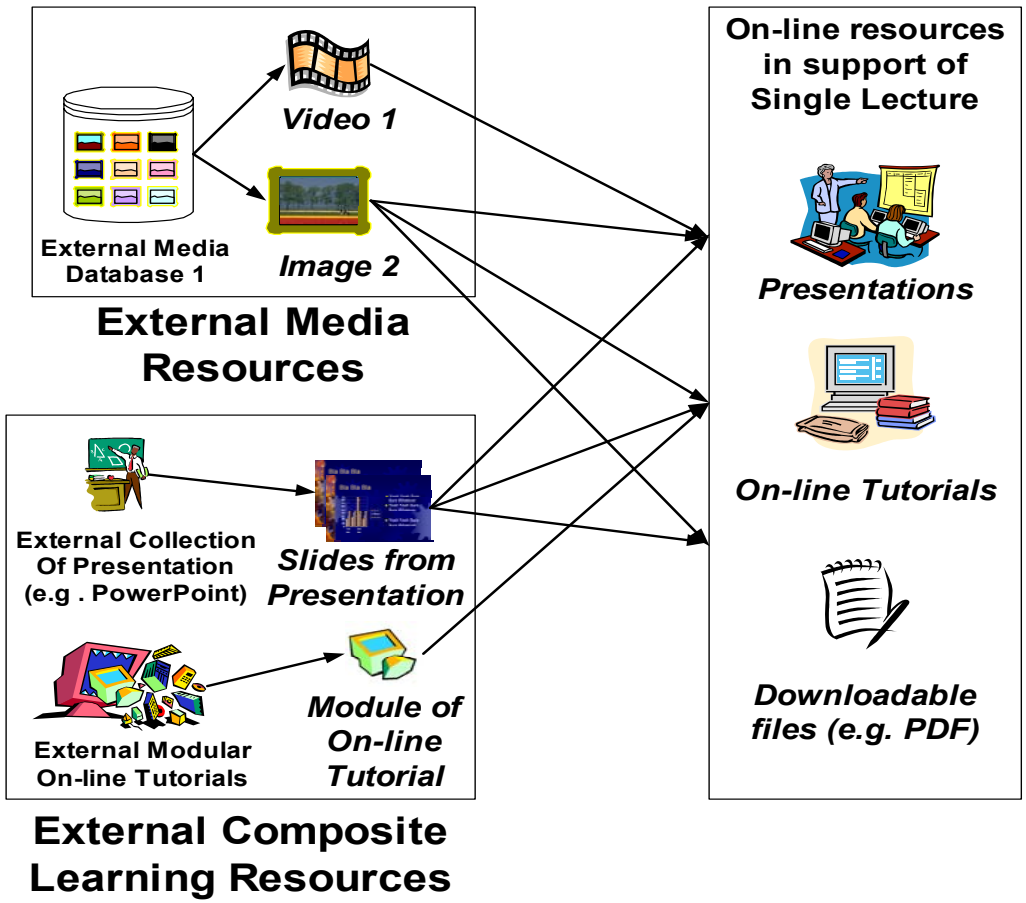
- In the UK… e.g. ARKive
  - 'Multimedia profiles' of Animal, Plant & Fungi + Habitats
  - Phase 1 – 30,000 images, 900hrs video, audio, maps, text…
  - Funding - UK Heritage Lottery Fund (£1.6m), New Opportunities Fund (£0.5m), HP Labs ($2m for technology research), …
- Others… SCRAN, British Pathe… other NOF projects, Commercial Organisations and NGOs…
- ARKive-ERA (Educational Repurposing of Assets) Project – Investigating key issues in designing systems to provide diverse users with appropriate access to multimedia assets…
- Funded by HP Labs as in support of work with ARKive

Repurposing

- Maximising use across range of user groups and individuals. Across types of media, platform, context…



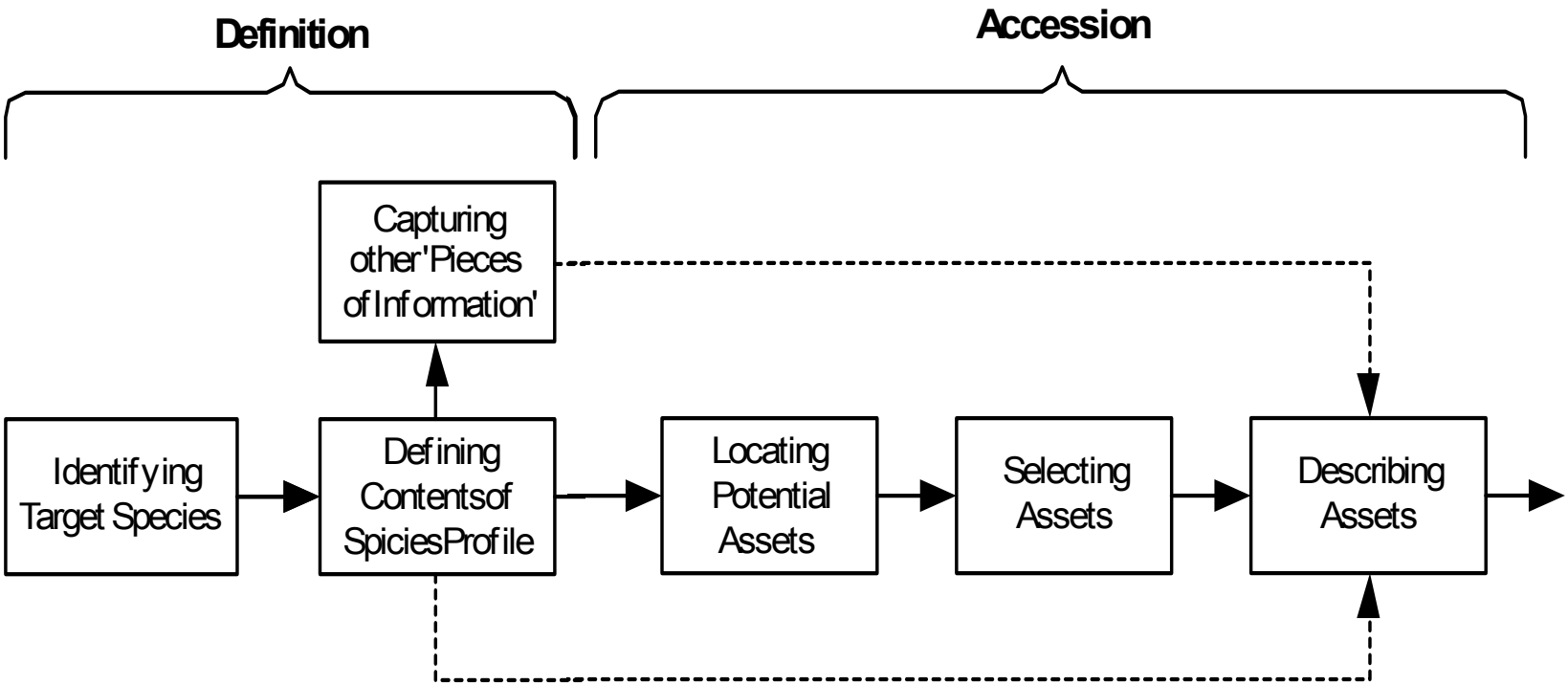
Repurposing



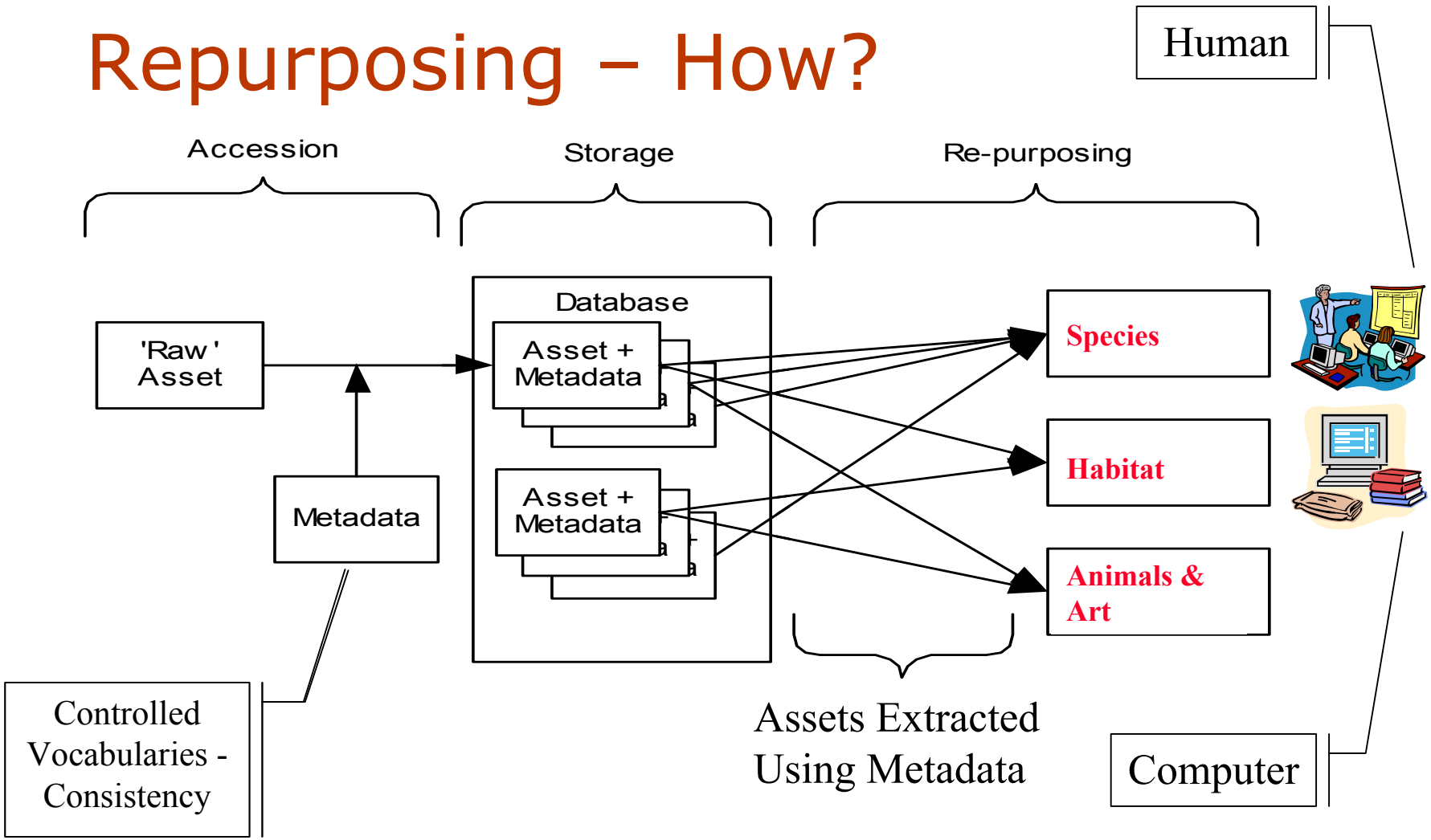
- Pulling together resources from different sources…

The ‘best’ resources

Repurposing How?
ARKive - Defining The Collection?



Repurposing – How?



Some Problems & Issues

- Capturing ALL the relevant information
  - Comprehensive description of species and their habitats
- Describing Resources
  - Choosing and creating core ‘vocabularies’
  - Inter-indexer consistency & assistance
- Capturing ‘pieces of information’
  - Capture and indexing of ‘pieces of information’
  - Representing controversial, partial and changing knowledge
- Finding Resources
  - Example: The Flying Bird Problem
- Providing access to diverse audiences
  - Examples: The hat problem & A fundamental dilemma
- Interoperation with internal & external systems…

Capturing ALL the Information

- Wanting to capture comprehensive multimedia profile of a species – appearance, life cycle, behaviours, inter-specific relationships…
  - different types of profile for different species e.g. plants and animals, and insects and mammals…
- Requires a means of representing the content of those profiles…
  1) To check that there are materials for ALL of elements of profile
  2) To check what is missing – what is still to be obtained
- But new information/knowledge is always being discovered →

Capturing 'Pieces of Information'

- During the research phase (which is always!) information is gathered about a species…
- These new 'pieces of information' specific to species need to be captured and integrated into the profiles → continuous revision of profile [& new concepts!]
- New 'pieces of information' about other things e.g. key scientists, geographic regions, how different groups classify types of flight etc… needs to be captured and made available i.e. indexed… so can be retrieved for repurposing/authoring new content
- And … Much knowledge is controversial, partial and changes…

Describing Multimedia Resources

- Terms for controlled vocabulary for describing appearance, life cycle, behaviours, inter-specific relationships… e.g. Mapping terms from one *biological* sub-discipline to another…
- Other subject disciplines & perspectives??? → next slide
- Inter-Indexer Consistency – 46% consistency in tagging moving images…
- Of course not only descriptive terms but many other aspects of multimedia to describe too – technical, administrative, preservation… using 'metadata standards'

Providing Access to Diverse Audiences

- The Hat Problem
- Diverse (esp. near orthogonal) perspectives on any single asset → specialist vocabularies e.g.
  - The Medical Subject Headings (MeSH) - there are more than 19,000 main headings
  - The Art & Architecture Thesaurus (AAT) - contains about 120,000 terms covering objects, textual materials, images, architecture and material culture from antiquity to the present.
  - UK National Curriculum (for schools) Metadata Schema contains some 2000 'subject keywords'
- Comprehensive Description? → Costs? Time? Expertise? Which to choose?

A Fundamental Dilemma

- Developers (e.g. ARKive) want to enable all potential educational users to gain maximum benefit from the assets held in their database
- They do not want to dictate or second guess how people might use an asset
- However: in order to allow anyone to find anything…
  - in a usable and effective manner
  - with finite time and budgets, to index assets
  - a limited choice of metadata terms must be made…
  - which in turn requires assumptions about the users and likely uses to which the resource might be put!

You don't want to (and can't) predict what your users will want to use the 'raw' multimedia assets for, but if you don't, your users can't get to the assets.

Finding Resources

- Both internal and external re-purposing require that assets/resources can be found at (re)authoring time →
- If you can't find a resource when you need it, is it really there? Effectively No.
- When traditional approaches break down e.g.
  - Cross searching across domains with little conceptual overlap
  - Example: Do birds that fly, fly?

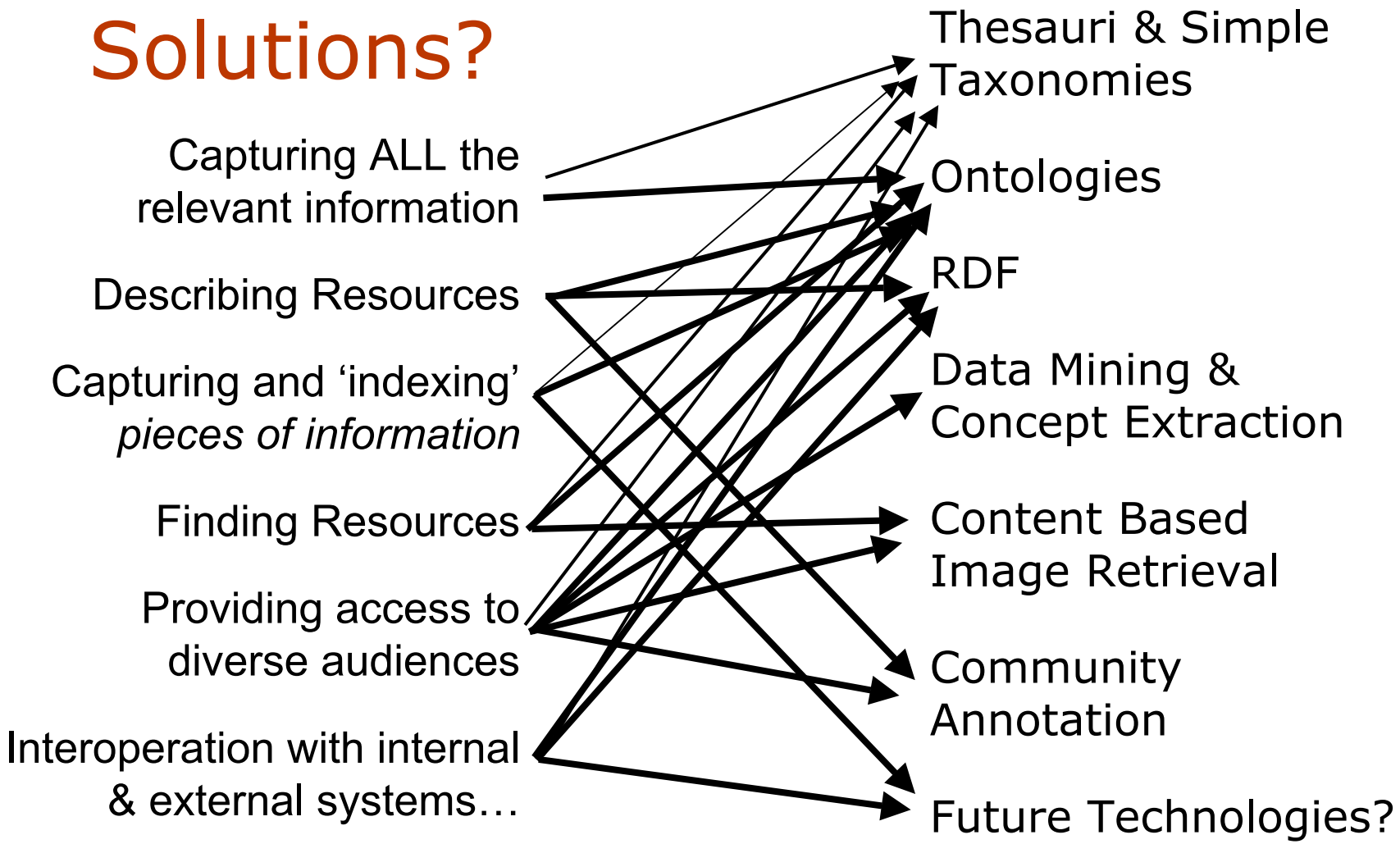
Interoperability

- BIG VALUE ADDED!!! If linked to other internal and external datasets e.g. news content, geographic, conservation organisations, Web cams…
- Biodiversity Domain
  - Few standard vocabularies and much controversy…
  - Diverse groups developing standards about related areas e.g. conservation, government, …
  - Politics & Politics are major issues
  - Human Knowledge is just really messy!

Solutions – Overview/Key Issues?

- Metadata and knowledge representation
- Getting into data sets with minimal conceptual overlap → Semantic Bootstrapping
- Auto or assisted/semi-automatic indexing
- Comprehensive cross subject domain mark-up and/or retrieval of resources
- Metadata standards and standards mappings

Solutions?


Capturing ALL the relevant information
Describing Resources
Capturing and 'indexing' *pieces of information*
Finding Resources
Providing access to diverse audiences
Interoperation with internal & external systems…
Thesauri & Simple Taxonomies
Ontologies
RDF
Data Mining & Concept Extraction
Content Based Image Retrieval
Community Annotation
Future Technologies?

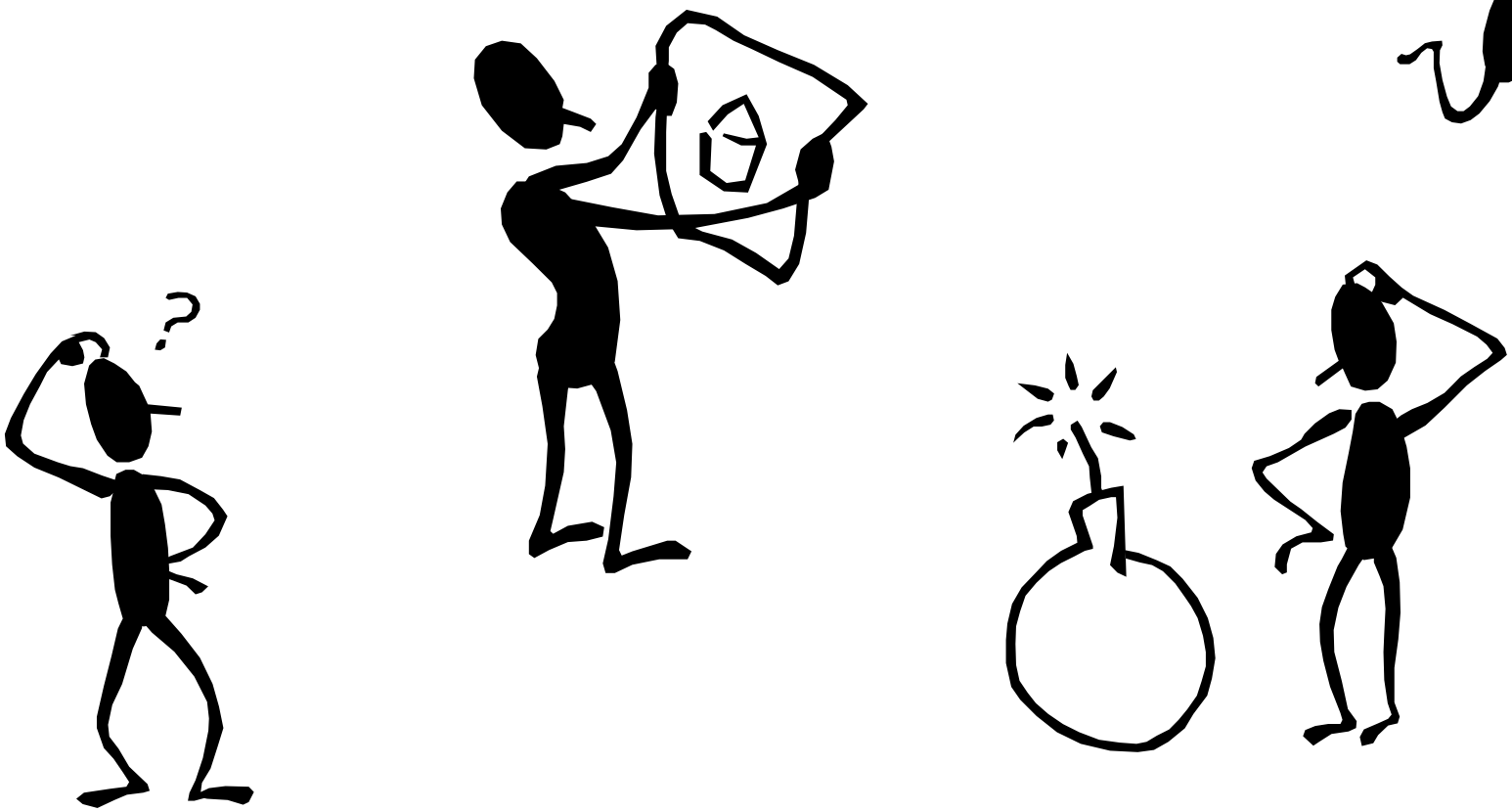
Solutions – Some Limitations and Issues

- Many seemingly simple things still difficult to represent [in RDF/ontology languages] e.g. ‘all birds, except x, y & z, fly.
- Many ‘hard’ problems e.g. controversial, partial and changing knowledge – human knowledge in most domains is messy!
- Ontologies are complex things and impose an ‘invisible’ world view on users – academic issues…
- Rapid and distributed ontology development & management is very problematic
- Ontology mapping still very problematic
- CBIR and concept extraction tools for non-text media not very good yet

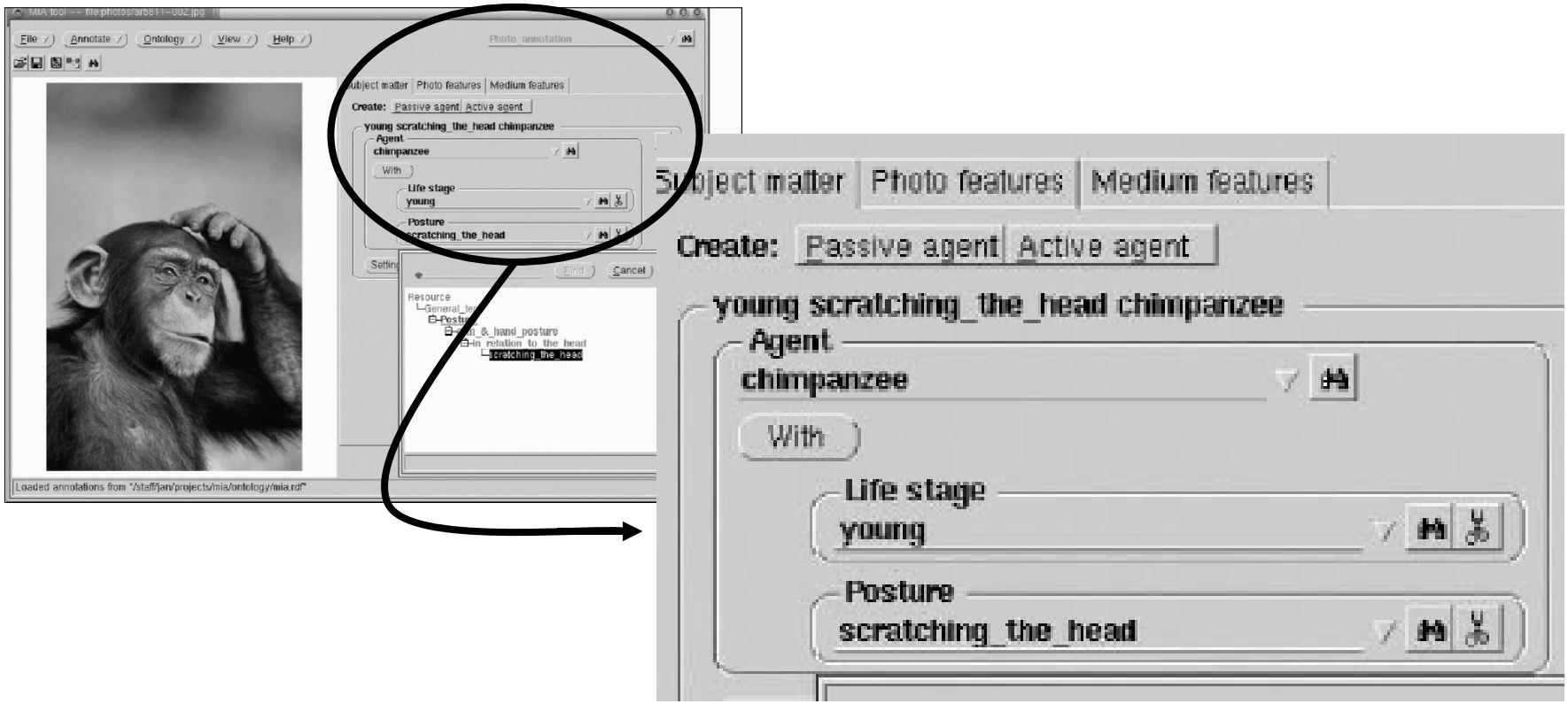
Conclusions

- ARKive was/is a very good example of a digitisation project – Majority of generic problems
- There were & are many problems! Some tractable with existing technologies others not yet. Some ‘simply’ to do with people
- The Semantic Web technologies and approaches offer part of a more comprehensive set of solutions to solve problems…

Questions



Machine Readable Ontologies



Schreiber et al (2001), Ontology-Based Photo Annotation, IEEE INTELLIGENT SYSTEMS, MAY/JUNE 2001

The Hat Problem

- Imagine that there is a database of 100,000 images of people in a wide variety of different settings
- The database is indexed using terms relating to identification of people (name, age,…) and event (time, place…)

- Now a milliner might see great potential for studying how people use hats.
- The database is likely to be a very useful resource, they could search to see how the style of hats has changed over time, or what types of hats are most popular, what percentage of women have bows on their hats? and many more specific questions
- However the database is not indexed using the concept of 'hat'…

BACK