CASPAR: Exploiting Infrastructures for preservation

Carlo Meghini
CNR ISTI
(slides by David Giaretta)

Drivers

- Fragility of digitally encoded information increasingly appreciated as a major concern
- This concern applies to almost every aspect of life
- Traditional memory institutions - limited viewpoint

- Action needed at many levels
  - **Community**
  - **Political/Funding**
  - **Technical**

Infrastructures for preservation

- There are various types of infrastructures which one needs to consider with respect to preservation
  - **Social / Legal / Financial / Organisational**
  - **Agreements / Standards**
  - **Technical components**

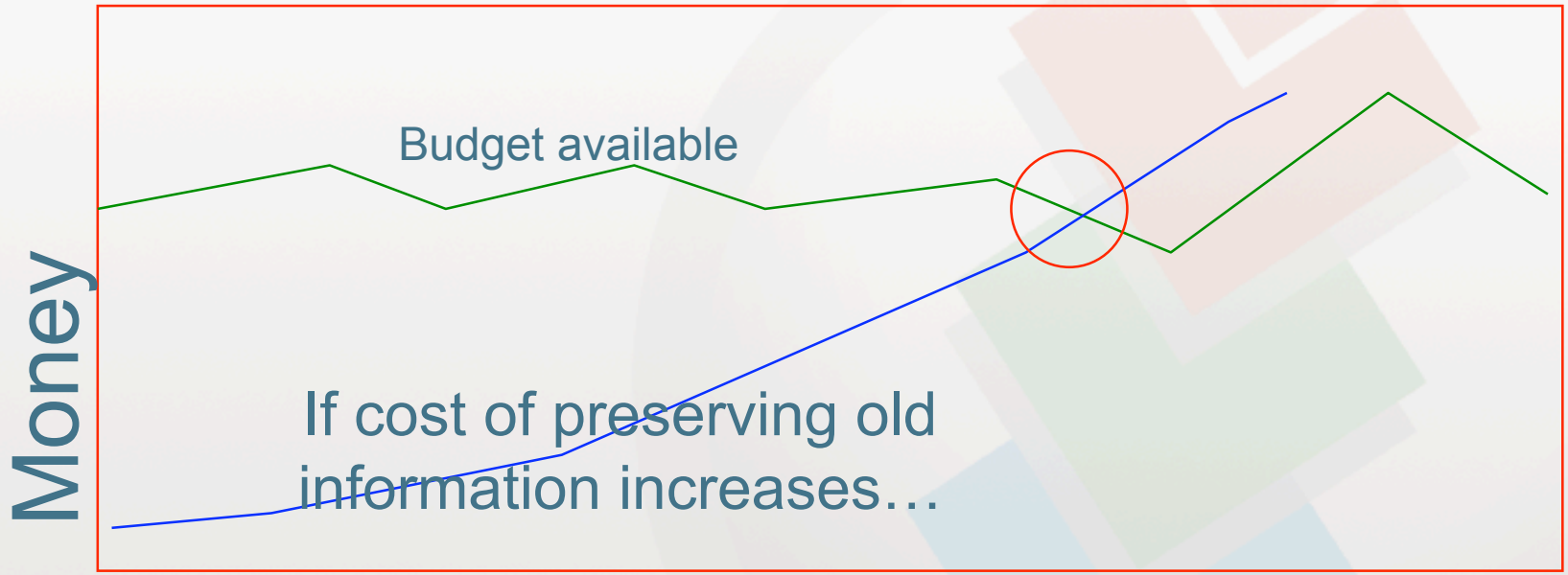
Digital Preservation

- Need to preserve information & **knowledge** – not just "the bits"
  - **Documents, videos are *rendered* – simple?**
  - **Data – must be processed – in new ways - harder**

Publication of data as well as documents

- Need to manage knowledge to keep archives alive through time
  - **Preservation is a process, not a one-time event**
  - **Preservation is expensive – costs need to be shared**
    - **The alternative is money – endless supplies of money**
  - **Open Archival Information Systems Reference Model (ISO 14721) provides a general conceptual framework and terminology**
    - *(http://public.ccsds.org/publications/archive/650x0b1.pdf)*
    - ***OPEN process – not just "Open Archives"***

Disincentives for preservation: cost



Need to show that costs are contained

Unfamiliar information

- Preservation
  - **Digitally encoded information which must be usable and understandable**
  - **Unfamiliar because of separation in time**
- e-Science/GRID/CyberInfrastructure for data
  - **Digitally encoded information which must be usable and understandable**
  - **Unfamiliar because of separation in discipline or location – even if created yesterday**

e-Infrastructures

- Doing preservation “right” produces gains in usage
  - **Preservation useful for e-science/GRID**
  - **Enables interoperability along several continua**
    - **Time**
    - **Discipline**
    - **Organisation**
- e-Infrastructures need components to support preservation – and gain interoperability
  - **Additional types of Finding Aids also useful**

Alliance for Permanent Access

- Part of strategy from Task Force for Permanent Access to the Records of Science
  - **With Research programme outline**
  - **Aim to "align" digital curation activities in the members**
- Members of Alliance:
  - **The European Science Foundation**
  - **European Space Agency**
  - **CERN**
  - **Max Planck Gesellschaft**
  - **Centre National d'Etudes Spatiales**
  - **Science and Technology Facilities Council**
  - **The British Library**
  - **Koninklijke Bibliotheek**
  - **Deutsche Nationalbibliothek**
  - **Joint Information Systems Committee**
  - **International Association of Scientific, Technical and Medical Publishers**
  - **National Archives of Sweden**
  - **Centre Informatique National de l'Enseignement Superieur**
  - **Digital Preservation Coalition**
  - **NESTOR**
  - **Perennisation des Informations Numeriques**
  - **Netherlands National Coalition for Digital Preservation**

http://www.alliancepermanentaccess.eu/

Key OAIS (ISO 14721) Concepts

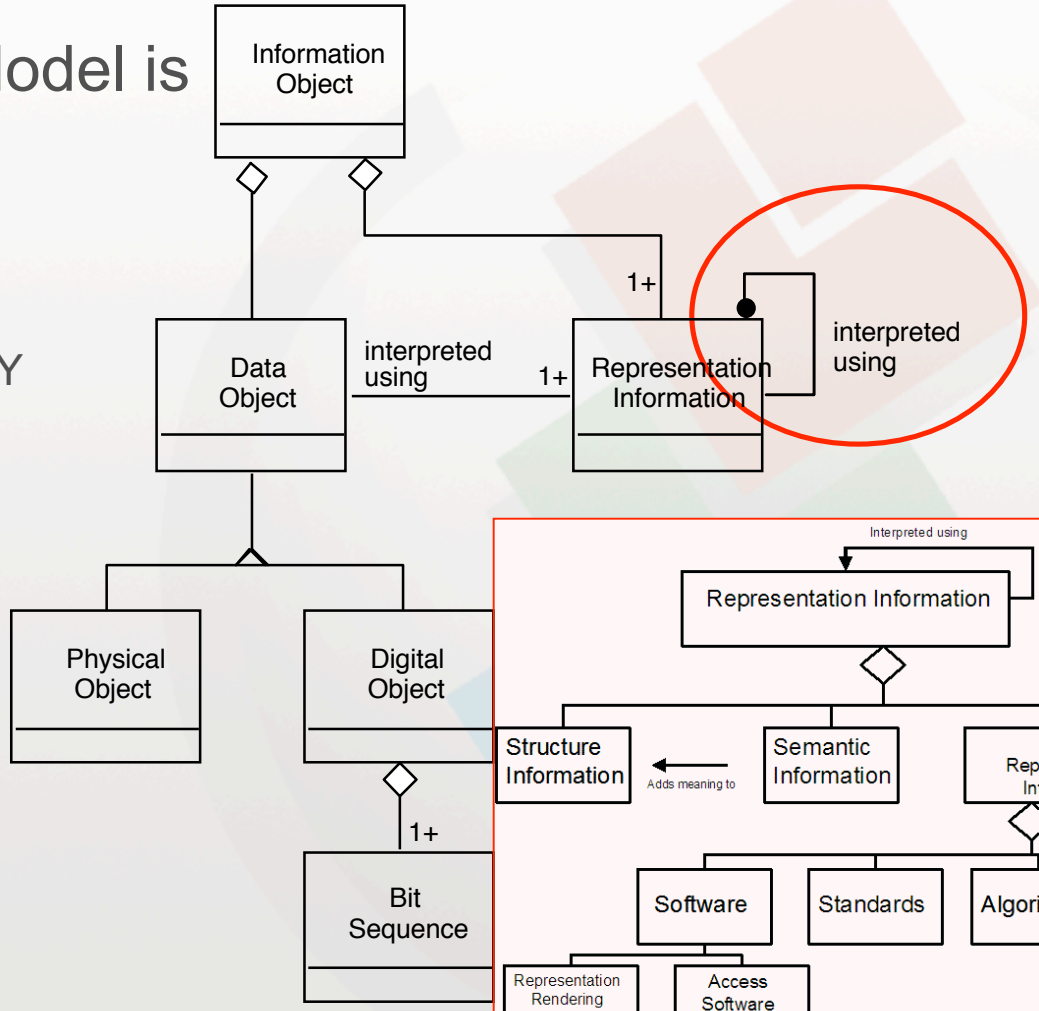
- Claiming “This is being preserved” is untestable
  - **Essentially meaningless**
- How can we make it testable?
  - **Claim to be able to continue to“do something” with it**
    - **Understand/use**
      - **Need Representation Information**
- Still meaningless…
  - **Things are too interrelated**
    - **Representation Information potentially unlimited**
  - **Designated Community**
- Plus many other concepts

Information Model & Representation Information

The Information Model is key

Recursion ends at KNOWLEDGE BASE of the DESIGNATED COMMUNITY

(this knowledge will change over time and space)



Information is the important thing

- What information?
  - **Documents……**
  - **Data…….**
- Original bits?
- Look and feel?
- Behaviour?
- Performance?
- Explicit/ Implicit/ Tacit

Information:

Any type of knowledge that can be exchanged. In an exchange, it is represented by data.

Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely.

Ensure that the information to be preserved is Independently Understandable to (and usable by) the Designated Community.

Sharing

- What can be shared?
  - **Care of the bits …..**
  - **Representation Information about….**
- How to share it?
  - **Need a way that is independent of domain and type**

Levels of applicability of OAIS

- Checklist:
  - **Representation Information (RepInfo)**
  - **Preservation Description Information**
  - **Packaging Information**
- Exhaustive text documentation
- RepInfo supporting:
  - **Interoperability**
  - **Automation**

Double payback: interoperability

- **Applicable in "GRID" context**
  - **usability now as well as later**

Audit & Certification

- Need for a way to judge digital repositories has been demanded for > 10 years
- RLG/NARA produced TRAC (Trusted Repository Audit Checklist)
- Follow on work on OAIS aims to lead to ISO standard on which an accreditation and certification process can be based
- Such a certification process could shape digital “market” in future

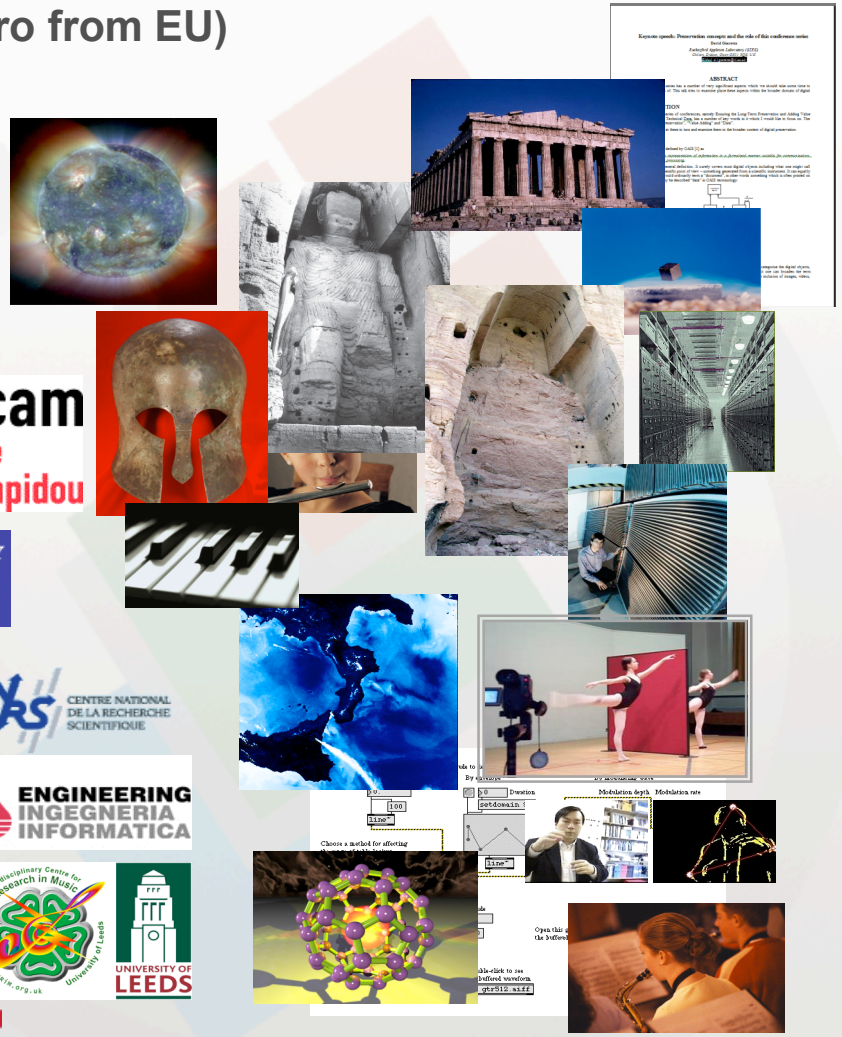
CASPAR Project

EU FP6 Integrated Project

Total spend approx. 16MEuro (8.8 MEuro from EU)

Started April 2006, for 42 months

David Giaretta is Co-ordinator



http://www.casparpreserves.eu

High Level Objectives

The guiding principle of CASPAR is the application of the OAIS Reference Model

- to research, develop and integrate advanced components to be used in a wide range of preservation activities.
- to create the CASPAR framework:
  - **the software platform that enables the building of services and applications that can be adapted to multiple areas**

The CASPAR consortium will seek to guarantee the future evolution of CASPAR. This ambitious goal will be pursued through:

- the building of the CASPAR preservation user community
  - **creating consensus around the initiative and gathering a critical mass of potential users**
- embedding the CASPAR framework and components within key memory organisations, both national and international.

CASPAR Aims

- **Produce tools and techniques to support digital preservation and make it easier to share the cost**
  - must **be relatively easy to use**
  - must **have a low "buy-in" in terms of effort required for adoption**
  - must **avoid requiring wholesale change of everyone else's systems**
  - must **be decentralised and reproducible so that it can live on after the formal end of the** CASPAR **project**
  - must **be** "preservable"
  - must **be open: open source, open standards**
- Cannot do everything
  - **Working closely with other projects**

Why CASPAR is special

- Digital preservation is hard
  - **OAIS view especially hard**
- No organisation/project can guarantee its own longevity
- Reduction of risk of losing information – cannot guarantee that nothing will be lost
  - **In the end it depends on money and interest**

- Need to be able to share the load
- Exploit OAIS Concepts to the fullest extent
- Broadly applicable
- Evidence of effectiveness

Test with science, cultural heritage and performing arts data

Sharing the burden of preservation

- Must be able to pass on custody of “the bits”
  - **May involve transformations over time**
  - **Can involve many duplicate copies**
- Need to harness the “wisdom of crowds”
  - **Wikipedia**
  - **BBC Wiki-radio – annotation**
  - **Google page-ranking**
  - **“given enough eyeballs, all bugs are shallow”**

BUT – if we are serious about the long term

- Need to be proactive
  - **To avoid death by neglect**

Infrastructure

- Somewhere to collect the contributed “wisdom”
  - **Registry/Repositories to store the additional information we need**
    - **separate from the archives which store the digital content - although a Registry/Repository is itself and archive**
- Something to remind people to take action
  - **Gap Manager**
  - **Orchestration Manager**
- More “local” tools
  - **Creation of Representation Information, PDI etc**
  - **Persistent Storage**
  - **…**

Things change/disappear

- Software
- Hardware
- Environment
  - **E.g. Network links to related information**
- People
  - **What is “common knowledge”**

How can we ensure that the information trapped in the “bits” remains understandable despite all these changes?

How can a digital curator even be aware of these changes?

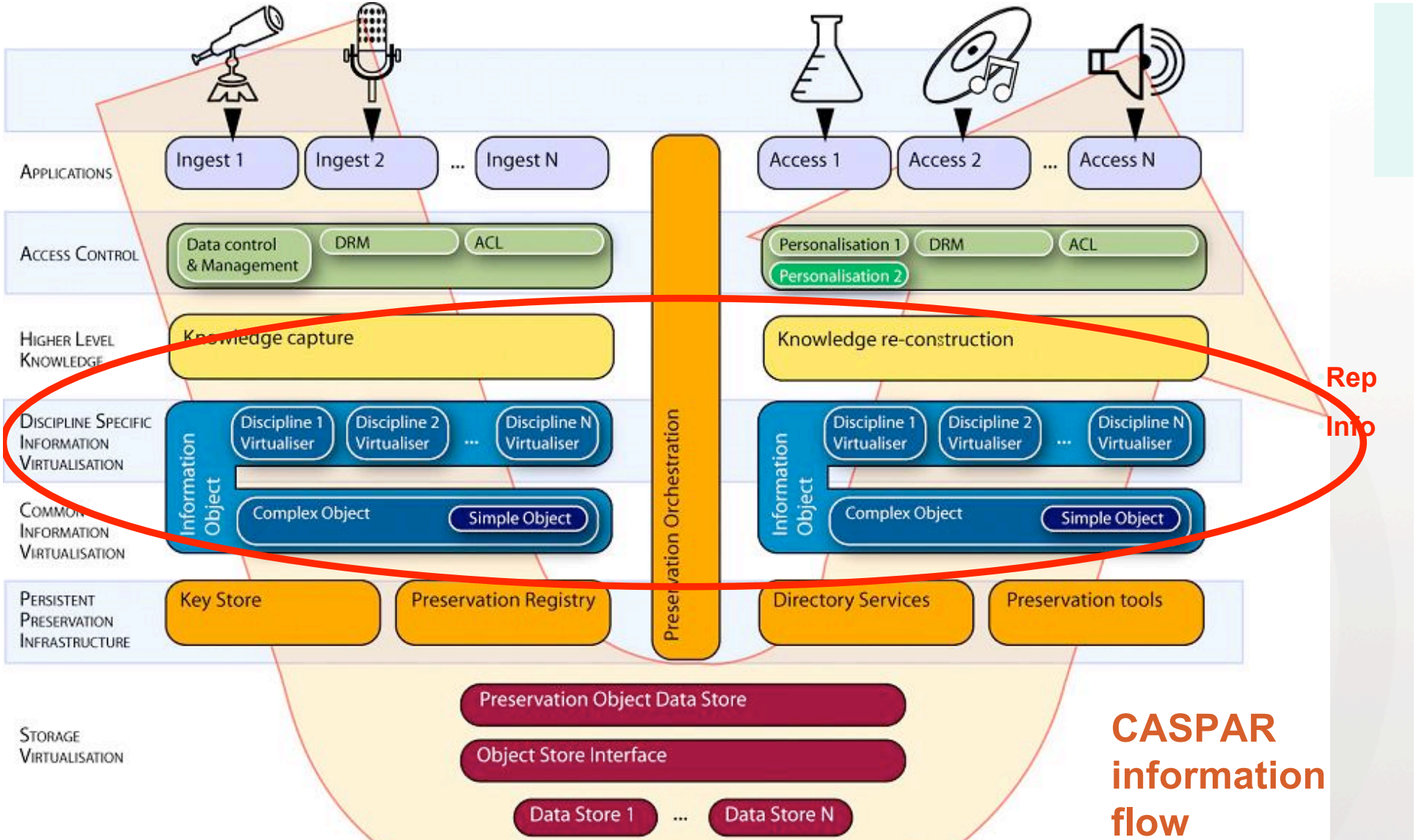
Validation

- How can we judge any proposed solution?

Live a long time

- CASPAR validation metrics:
  - **Theoretic underpinning**
  - **Testbed scenarios addressing real issues**
    - **No "hand-waving" – use what is there now**
    - **Accelerated lifetime tests**
      - **Hardware and Software**
      - **Environment**
      - **People**
  - **Improved "trustability"/"certifiability"**

Evidence - not proof



Rep

Info

CASPAR information flow architecture

How do we capture the Representation Information?

Preservability Infrastructure

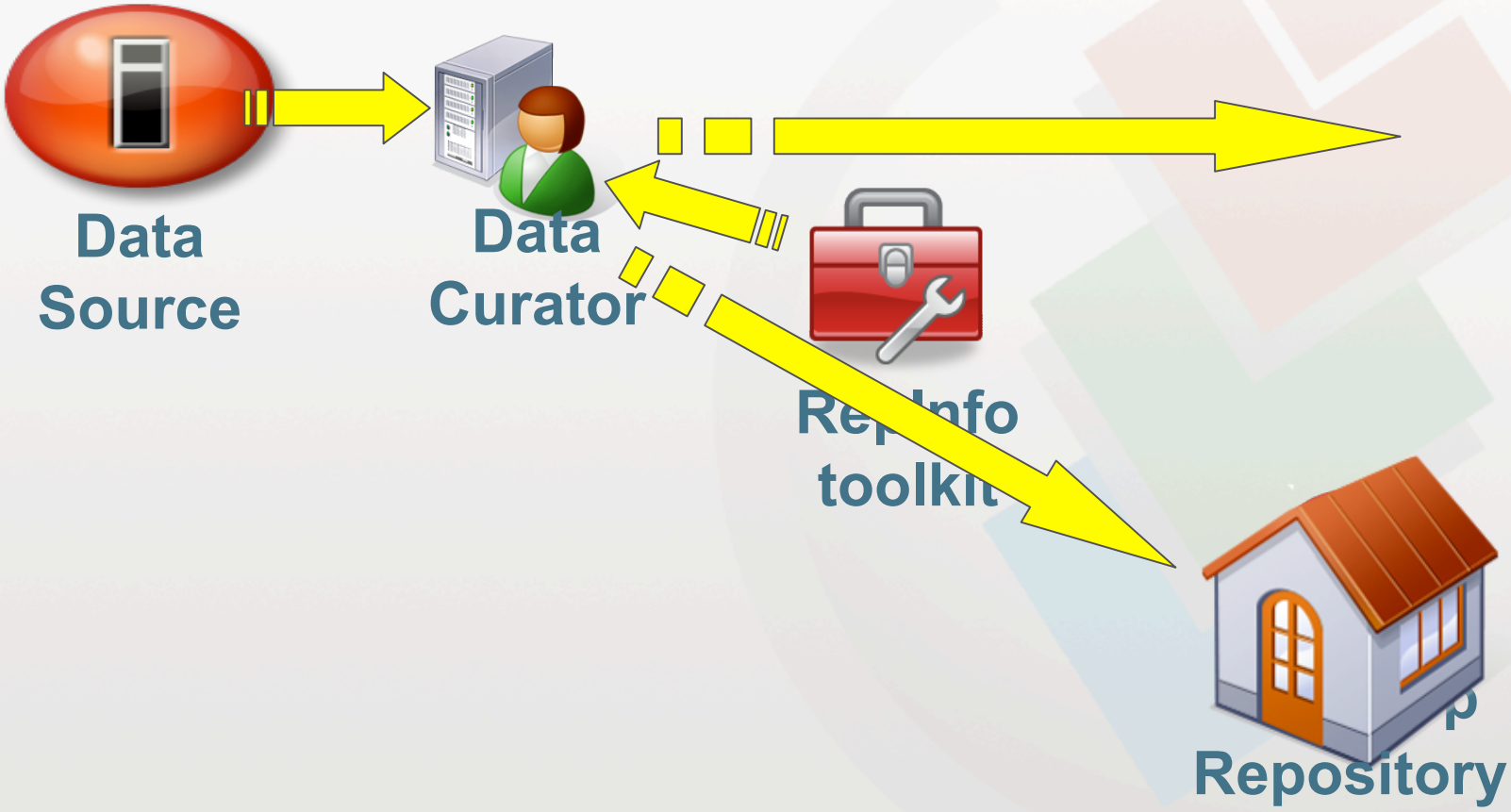
- Persistent Identifiers
- Distributed and Persistent Storage
- Identifying relevant information
- Registries of Representation Information
- Messaging infrastructure
- Workflow
- …….

END

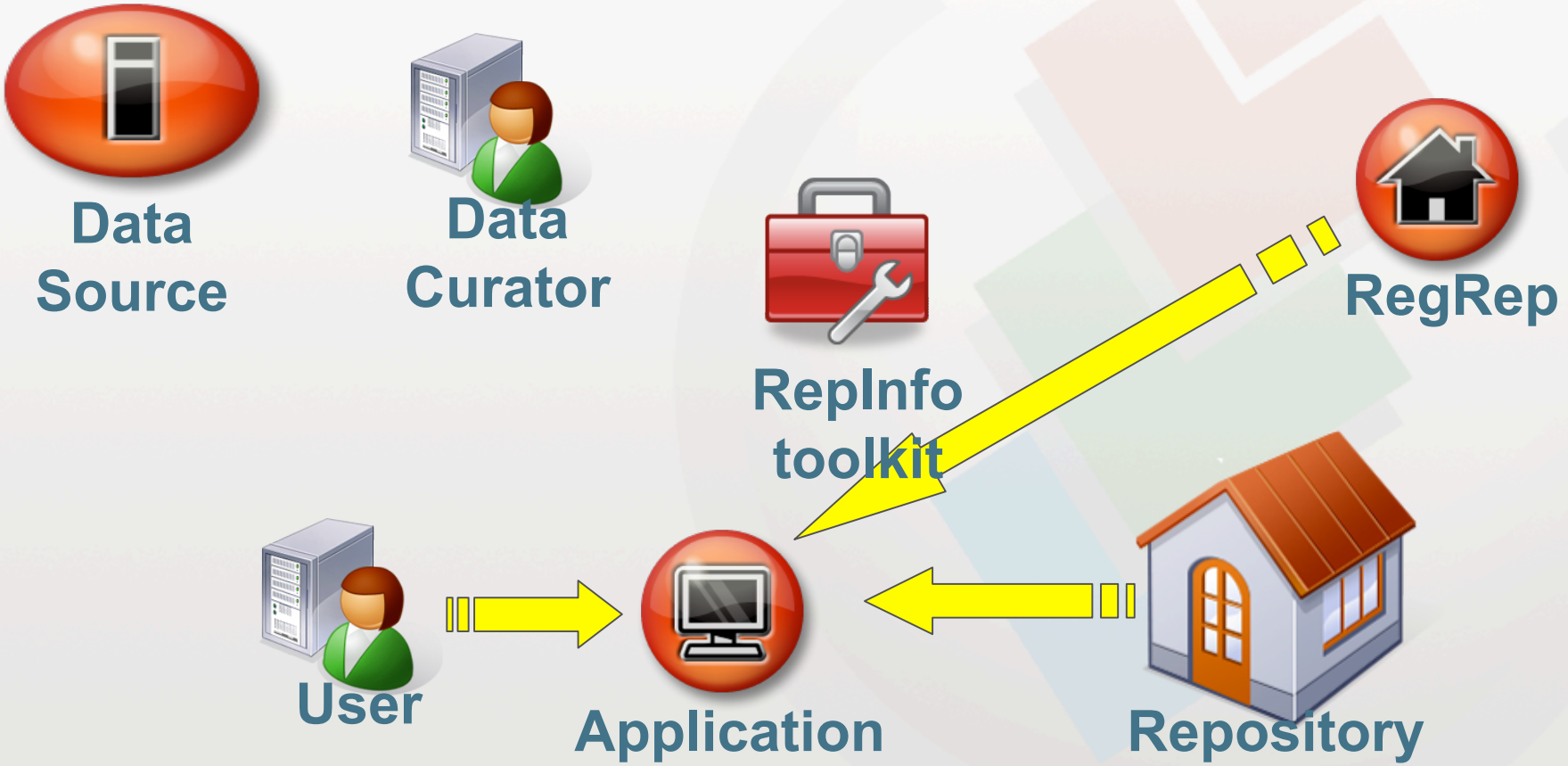
Infrastructures: Computation, Data, Information

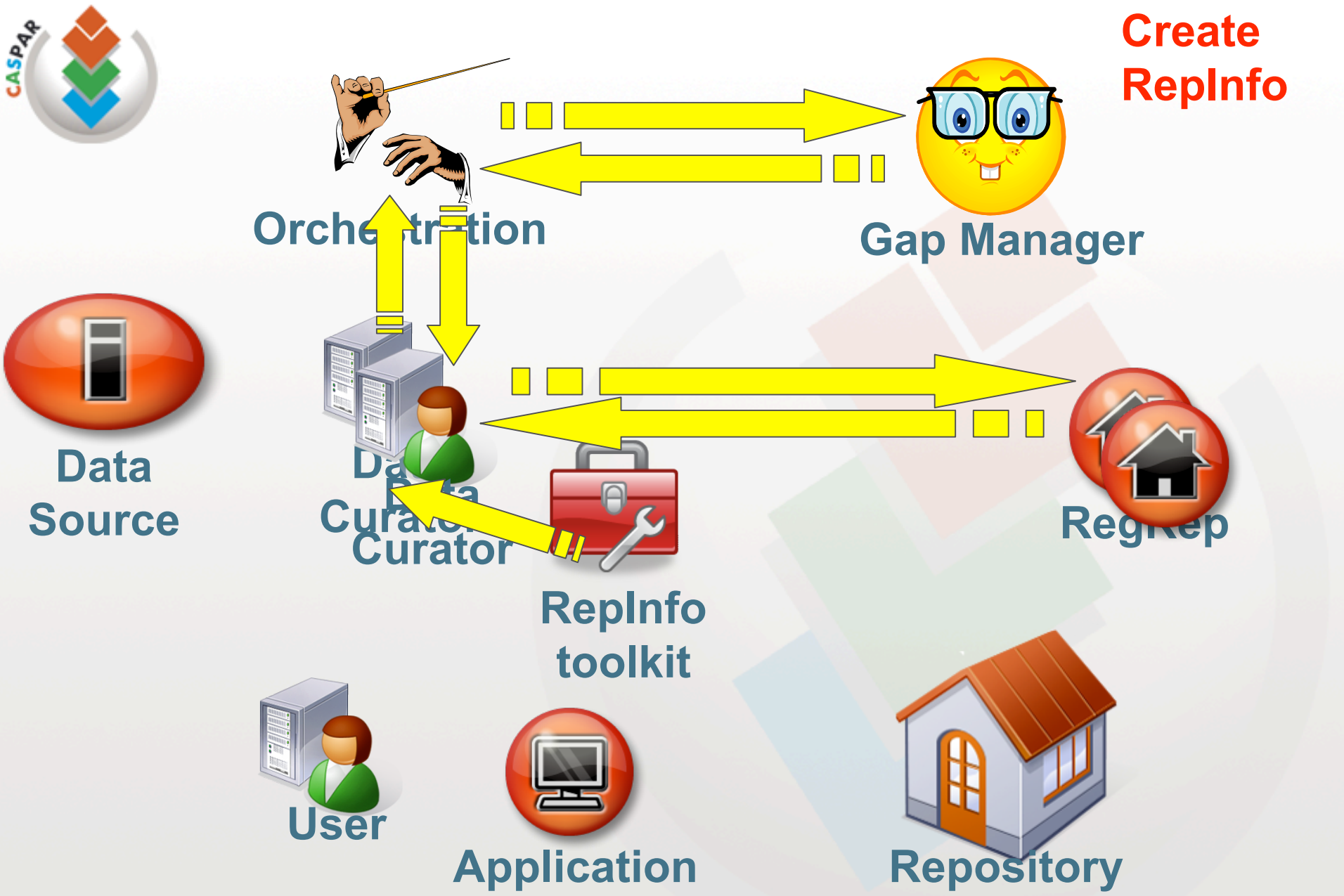
- Computation:
  - **BLOB transfers, pre-planned applications**
- Data:
  - **More structured BLOB transfers, selected applications**
- Information:
  - **Information transfers, appropriate processes applied**

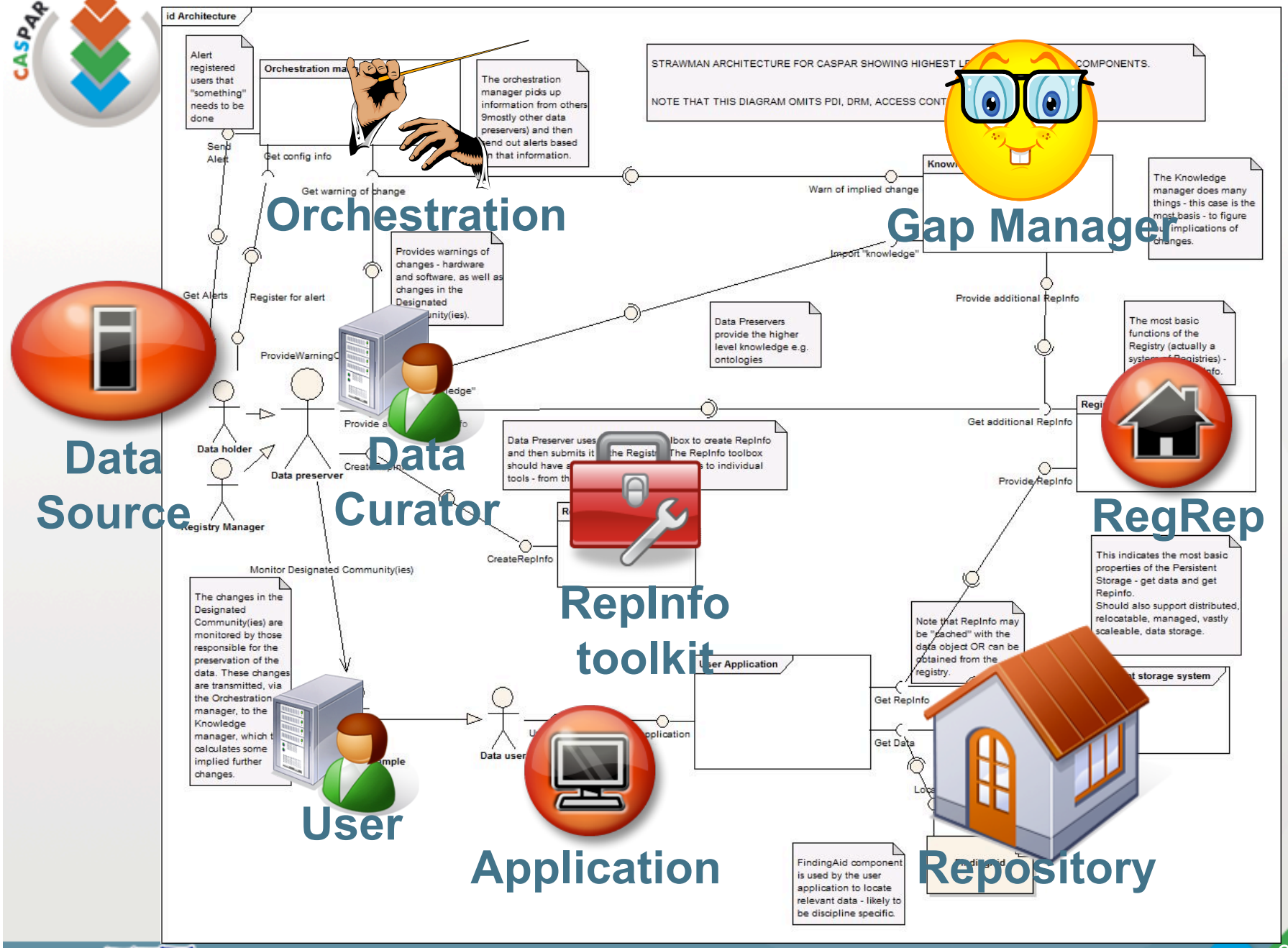
Create RepInfo



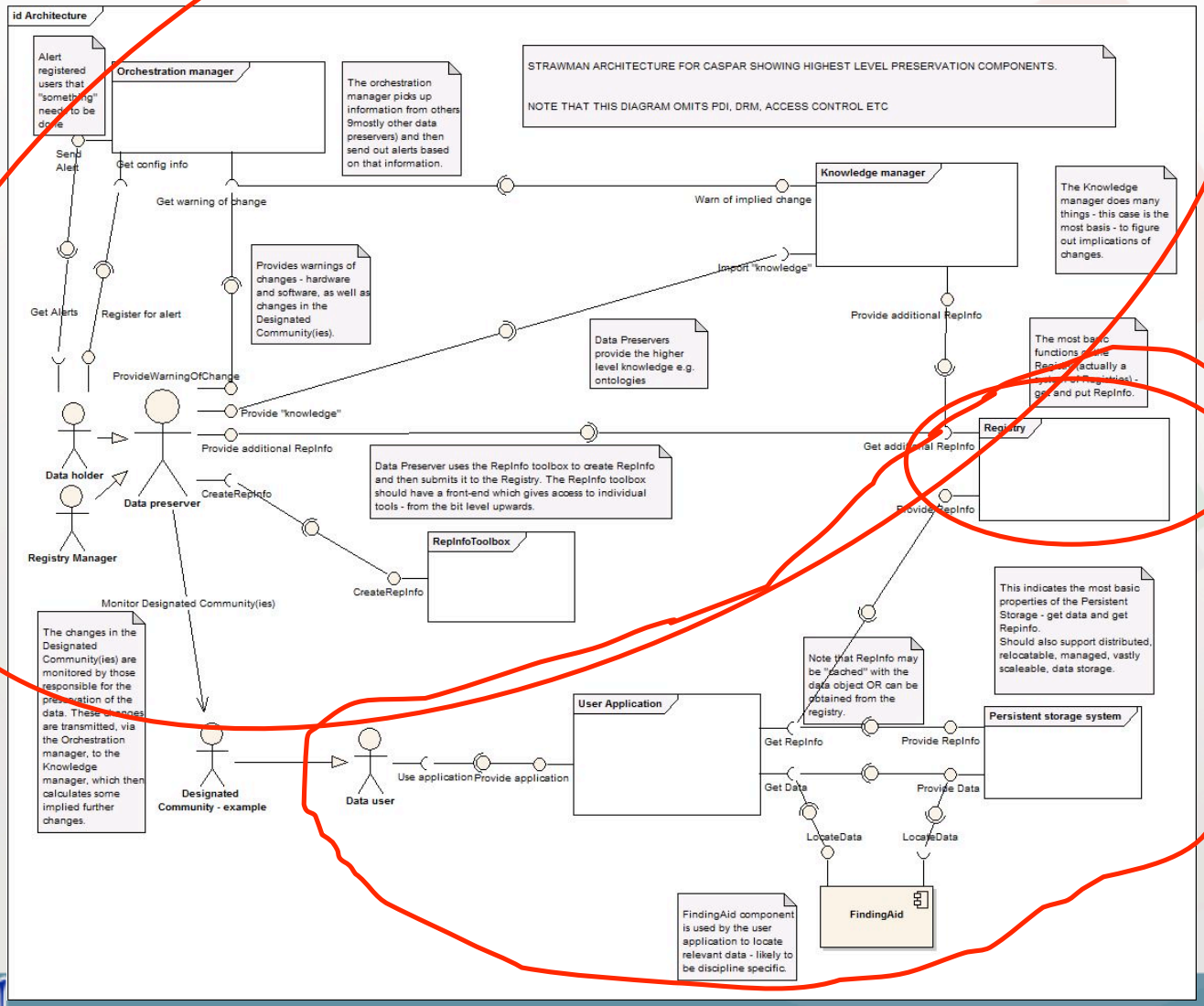
Use Data


Data Source
Data Curator
RepInfo toolkit
RegRep
User
Application
Repository


Create RepInfo
Orchestration
Gap Manager
Data Source
Curator
RegRep
RepInfo toolkit
User
Application
Repository


id Architecture
Alert registered users that "something" needs to be done
Orchestration ma
The orchestration manager picks up information from others 9mostly other data preservers) and then end out alerts based n that information.
STRAWMAN ARCHITECTURE FOR CASPAR SHOWING HIGHEST L COMPONENTS.
NOTE THAT THIS DIAGRAM OMITS PDI, DRM, ACCESS CONT
Send Alert
Get config info
Get warning of change
Orchestration
Warn of implied change
Know
Gap Manager
The Knowledge manager does many things - this case is the most basis - to figure u implications of changes.
Import "knowledge"
Provide additional RepInfo
Provides warnings of changes - hardware and software, as well as changes in the Designated nity(ies).
Get Alerts
Register for alert
Data Preservers provide the higher level knowledge e.g. ontologies
The most basic functions of the Registry (actually a syst Registries) - nfo.
ProvideWarningC
ledge"
Provide a fo
Data holder
Data preserver
Data Source
Data Curator
Get additional RepInfo
Regi
Data Preserver uses lbox to create RepInfo and then submits it he Regist The RepInfo toolbox should have a to individual tools - from th
Provide RepInfo
RegRep
CreateRepInfo
RepInfo toolkit
Registry Manager
Monitor Designated Community(ies)
This indicates the most basic properties of the Persistent Storage - get data and get RepInfo.
Should also support distributed, relocatable, managed, vastly scaleable, data storage.
The changes in the Designated Community(ies) are monitored by those responsible for the preservation of the data. These changes are transmitted, via the Orchestration manager, to the Knowledge manager, which t calculates some implied further changes.
Note that RepInfo may be "cached" with the data object OR can be obtained from the registry.
User Application
t storage system
Get RepInfo
Get Data
Data user
pplication
mple
User
Application
Repository
FindingAid component is used by the user application to locate relevant data - likely to be discipline specific.

Sharing the burden of preservation

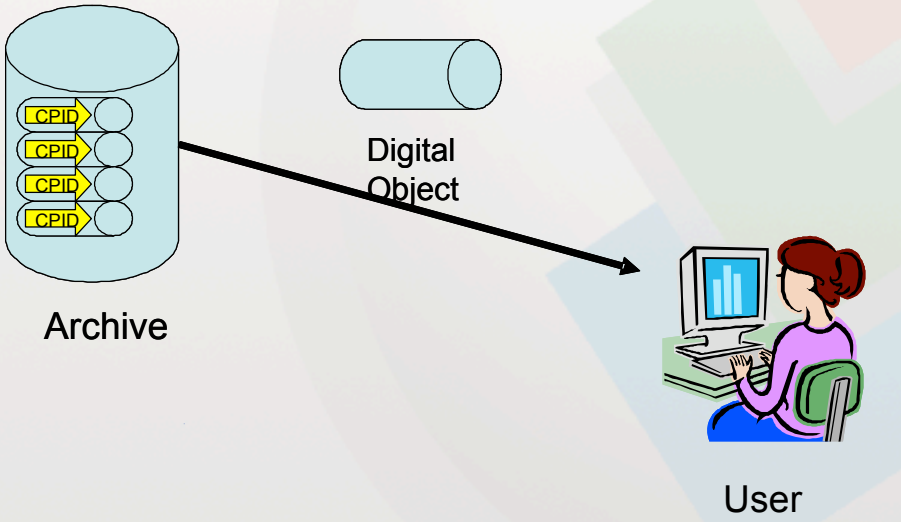


Registry/Repository Scenario

User gets data from archive.

User may be unfamiliar with the data so needs Representation Information. Some may be packaged with the data

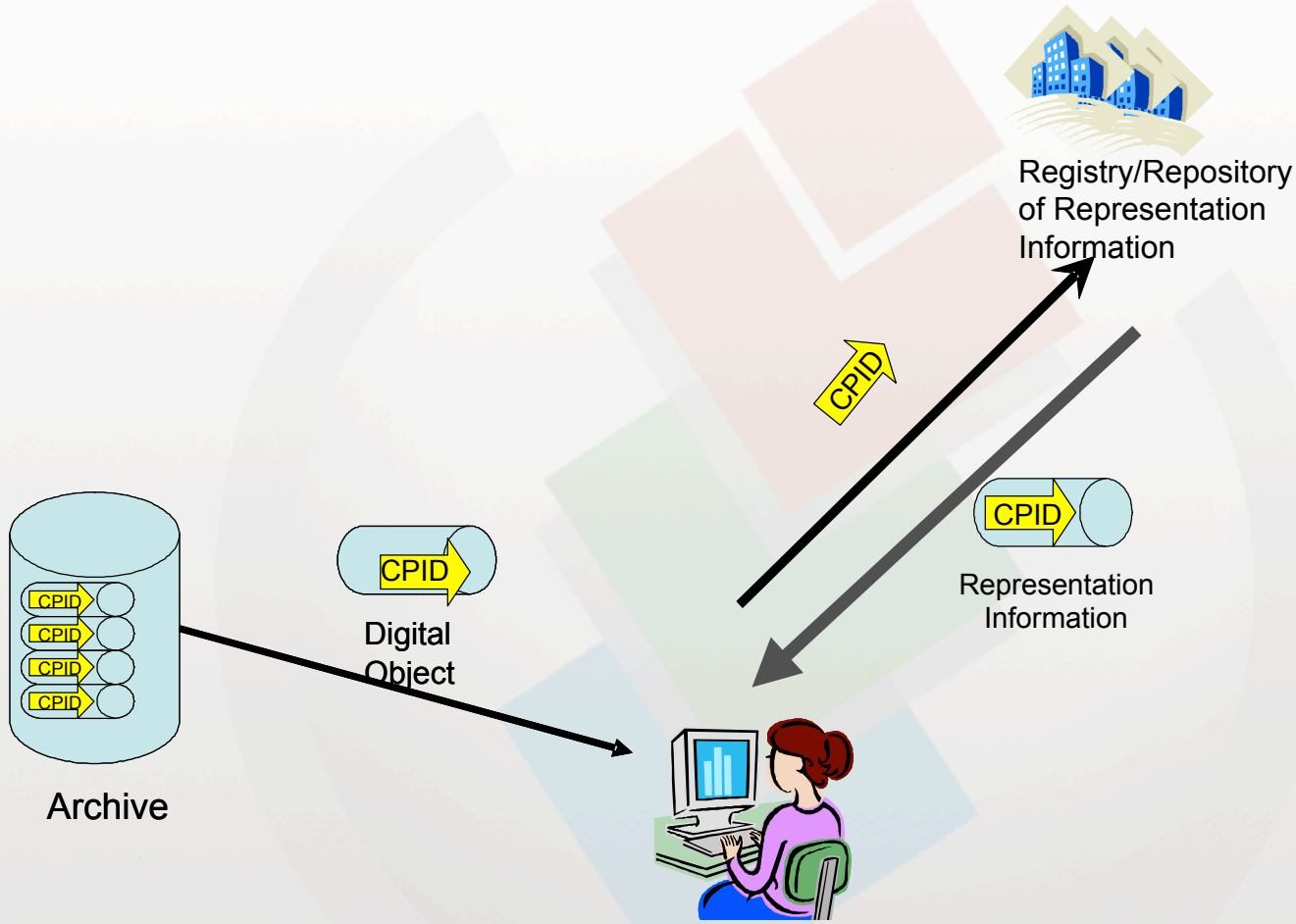
The Digital Object could have some RepInfo packed with it.



May need MORE Representation Information.

Obtain this from one or more Registry/Repository of Rep.Info.

Do not want to guess which RepInfo is needed – use an identifier (CPID) associated with the data



At some point in the future there is a need for further Representation Information.

RepInfo Toolkit

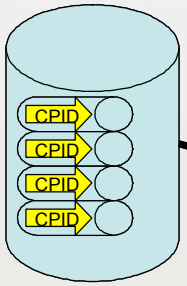
Registry/Repository of Representation Information

CPID

CPID

Representation Information

Someone creates the necessary RepInfo using whatever tools are appropriate.



Archive

CPID

Digital Object

Use of RepInfo