gCube Interoperability Framework

Leonardo Candela, George Kakaletris, Pasquale Pagano, Giorgos Papanikos, and Fabio Simeoni

2nd DL.org Workshop

Making Digital Libraries Interoperable:
Challenges and Approaches

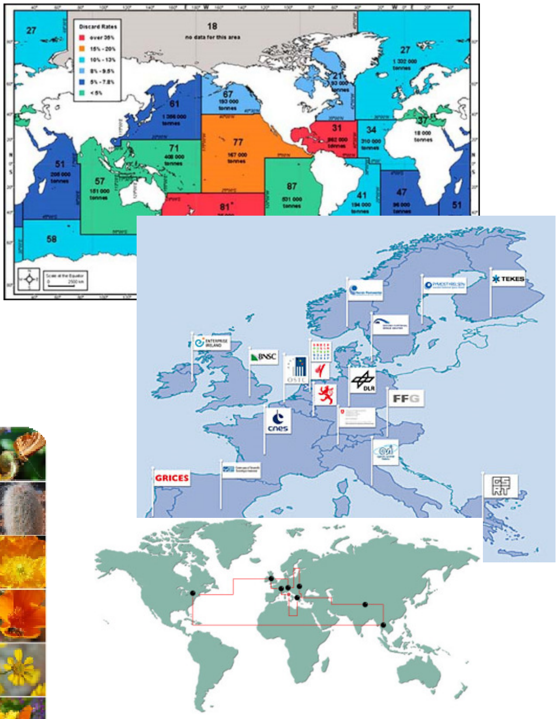
9-10 September 2010
Glasgow, Scotland, UK

Outline

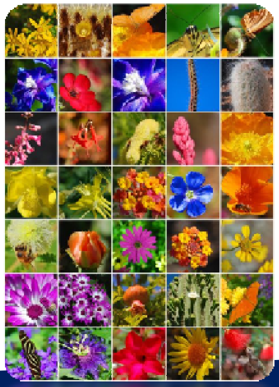
- Context
- Interoperability Approaches
  - Resource Discovery
  - Data Access
  - Data Discovery
  - Process Execution
- Success Stories
- Conclusions

Foundation

- gCube framework is the result of several years spent with wold-wide distributed international organizations
  - FAO, ESA, WorldFish Center, CERN, …



- We learnt how to
  - Appreciate the value of the differences
  - Work with such richnesses
  - Valorise the differences



Context: gCube as a Digital Library System

- A Digital Library System is a possibly distributed system that collects, manages and preserves for the long term rich digital content, and offers to its user communities specialised functionality on that content, of measurable quality and according to codified policies

[The Digital Library Reference Model]

The gCube data infrastructure enabling framework provides DL functionality by:

Federating exiting digital content

maintained in a variety of tailored repository systems

Supporting the generation of new digital content

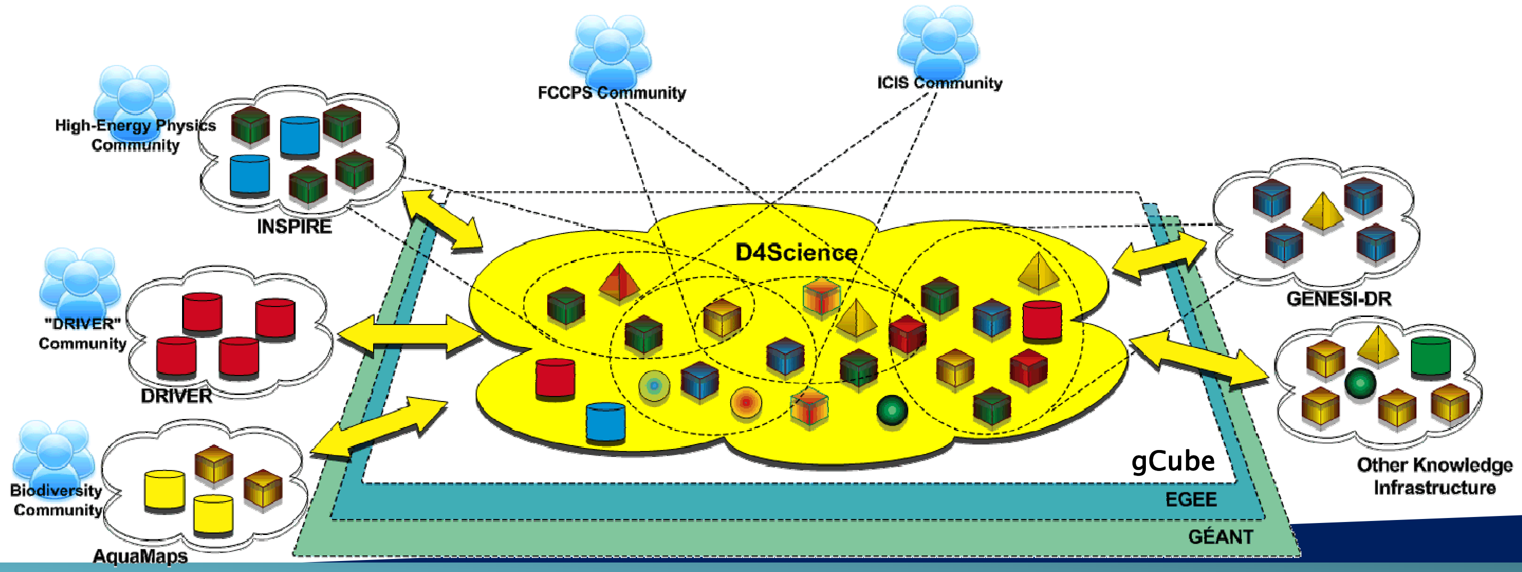
by exploiting heterogeneous computational platforms

Providing discovery and access capabilities

on diversely described and modeled digital content

Context: gCube as an e-Infrastructure ecosystem enabling framework

- By bridging a number of well-established systems and standards from various domains
  - including high-energy physics, biodiversity, fishery and aquaculture resources management
- gCube realises an e-Infrastructure ecosystem



Context: interoperability in gCube

- Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged

[IEEE Glossary]

- gCube defines interoperability as the ability of an arbitrary number of information systems to collaborate into achieving a common cause

Interoperability Approaches

Approaches and solutions to achieve interoperability in gCube:

- Blackboard-based
  - asynchronous communication between components in a system
  - one protocol to R/W and one language to specify messages
- Wrapper/ Mediator-based
  - translates one interface for a component into a compatible interface
- Proxy-based
  - exposes the same interface but allows additional operation over received calls
- Adaptor-based
  - provides a unified interface to a set of other components interfaces and encapsulates how this set of objects interact
- Broker-based
  - Specialises an Adaptor by coordinating communication

Interoperability Approaches: Resource Discovery

- Each resource is represented by a profile (metadata) characterising:
  - the interface
  - the state
  - the list the dependencies
  - the run-time status
  - the policies
  - the configuration
  - the pending tasks to execute

- A Resource profile
  - is published by the resource owner
  - is discovered by the resource consumers asynchronously through a common resource-independent protocol

- gCube offers a distributed and scalable Information System (blackboard) to store, discover, and access resource profiles

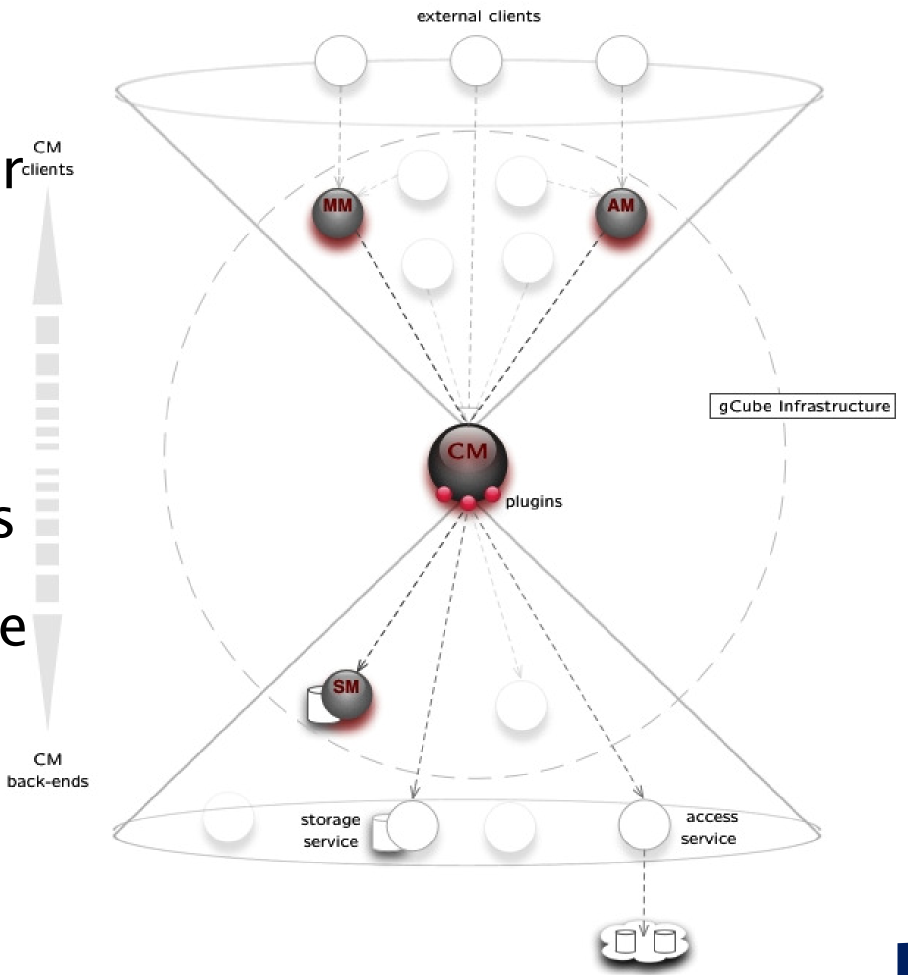
Interoperability Approaches: Data Access [1/4]

- Data Access and Discovery interoperability solution relies on the gCube Open Content Management Architecture (OCMA) design patterns:
  - terminology
    - *repository*: variety of back-ends including storage services as well as access services to content stored further afield
    - *model*: various media and structure
    - *protocol*: various access APIs

    “access type”
  - assumptions
    - content is created, accessed, and distributed in *documents*
    - documents are aggregated in *collections*
    - collections are hosted in *repositories*
  - goals
    - *embrace heterogeneity*: support diverse locations, model, protocols
    - *hide heterogeneity*: abstract over diverse locations, protocols, models
    - *scale*: retain good throughput under heavy load

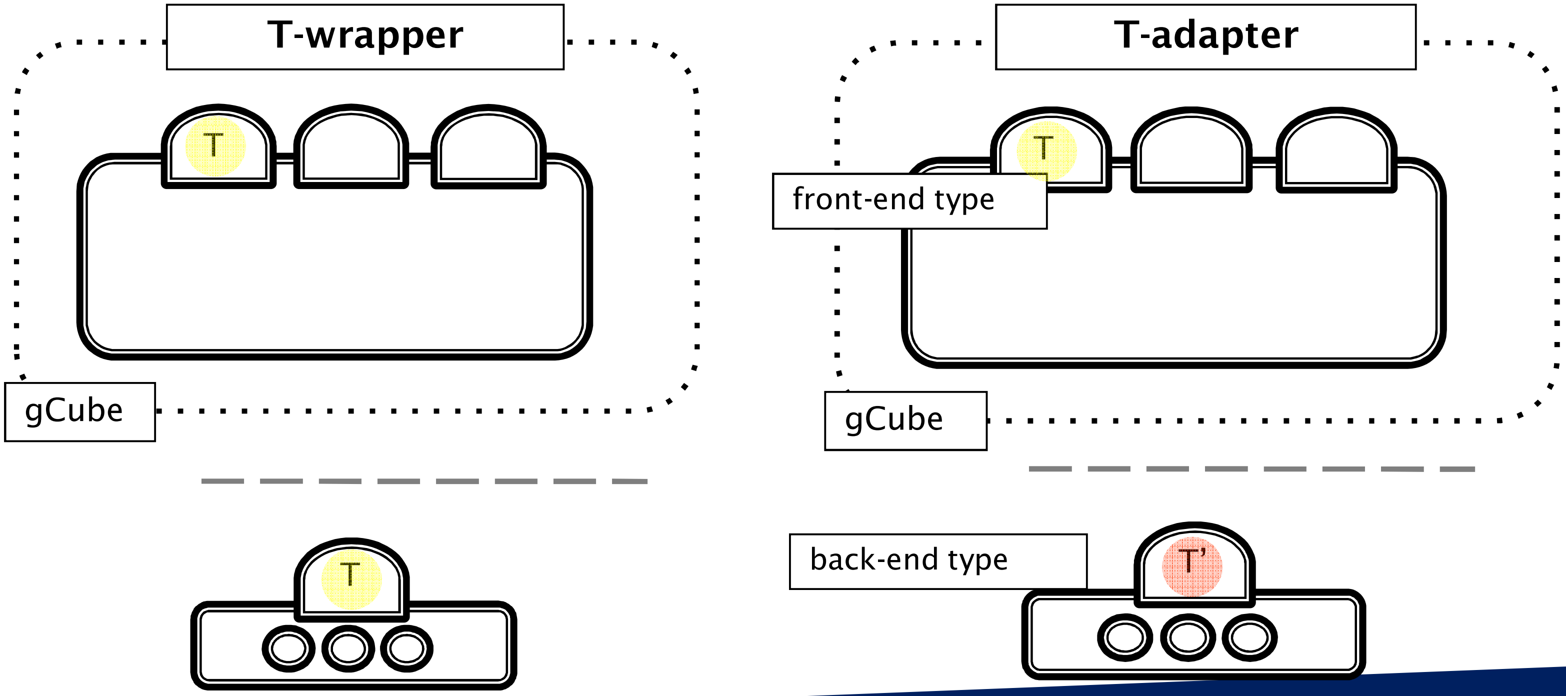
Interoperability Approaches: Data Access [2/4]

- Hourglass Topology Architecture
  - Central is the Content Manager
    - provides uniform access to heterogeneous content
  - CM relies on plugins to dynamically adapt to an arbitrary number of back-ends
  - Back-ends may include storage services as well as access services
  - CM models content as edge-labelled trees



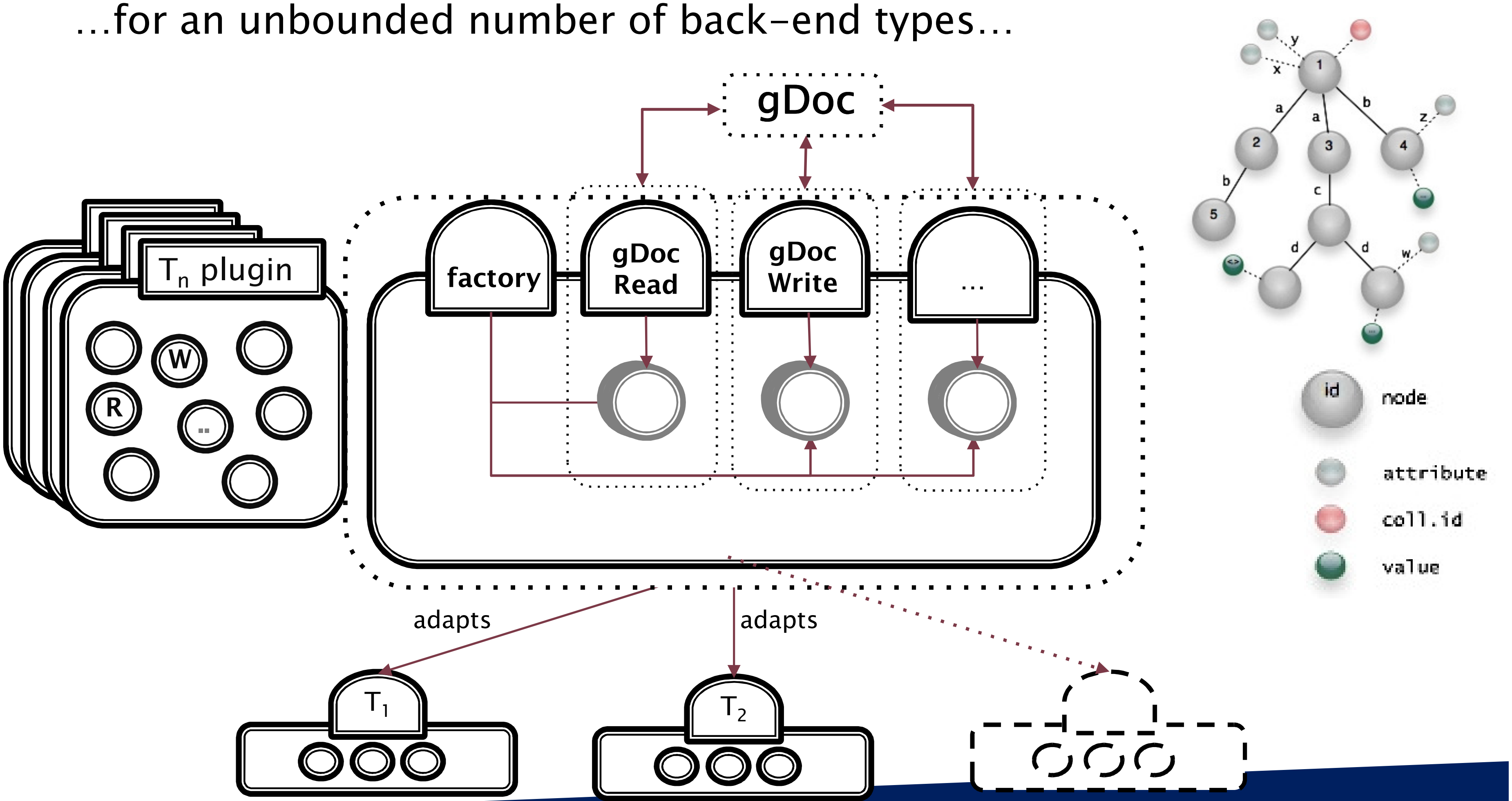
Interoperability Approaches: Data Access [3/4]

- A Repository R may already offer a native T-interface to a collection C. In this case the **CM acts as a wrapper for R.**
- A Repository R may offer a different T'-interface to a collection C. In this case the **CM acts as an adapter for R.**



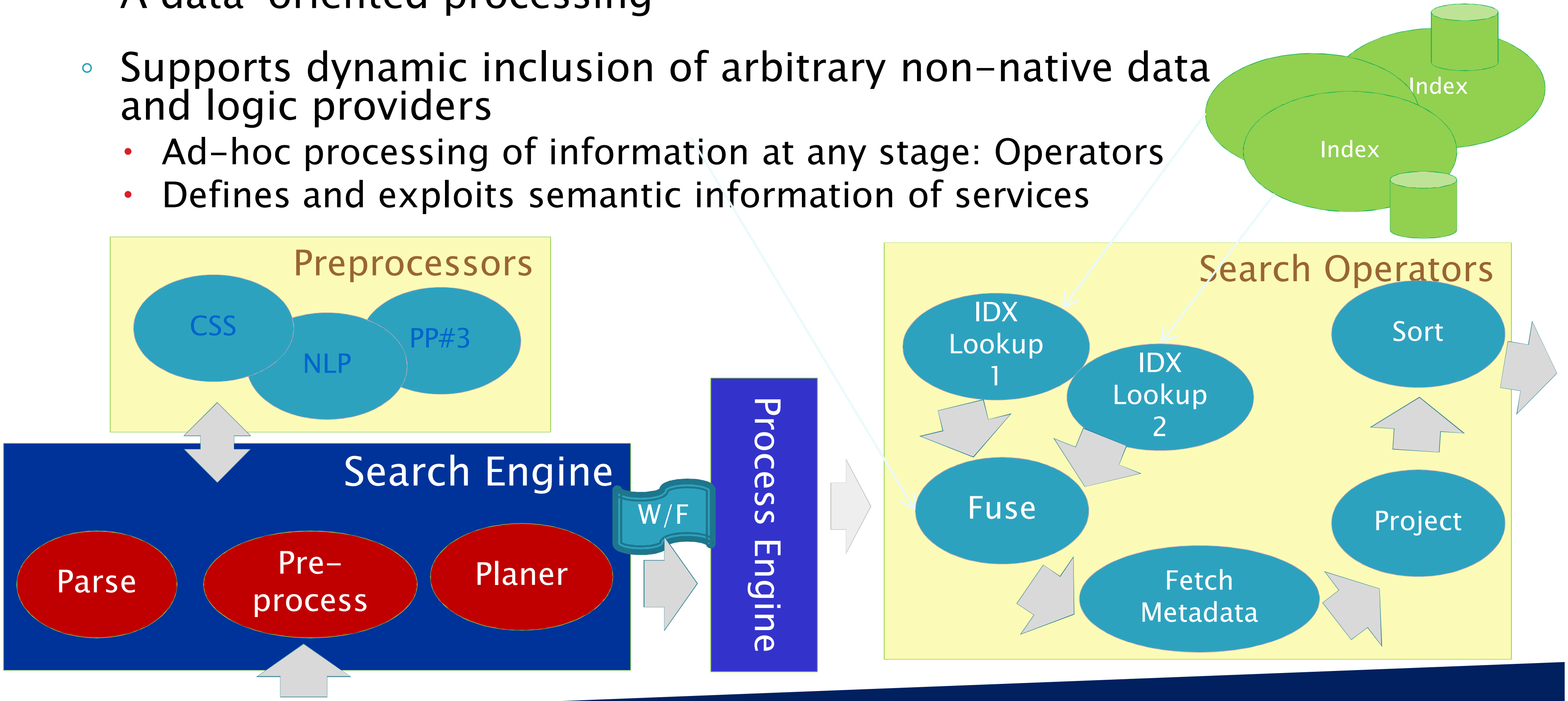
Interoperability Approaches: Data Access [4/4]

…for an unbounded number of back-end types…



Interoperability Approaches: Data Discovery

- The gCube Data Discovery interoperability framework is based on an open-ended set of preprocessor and operator mediators
  - A data-oriented processing
  - Supports dynamic inclusion of arbitrary non-native data and logic providers
    - Ad-hoc processing of information at any stage: Operators
    - Defines and exploits semantic information of services

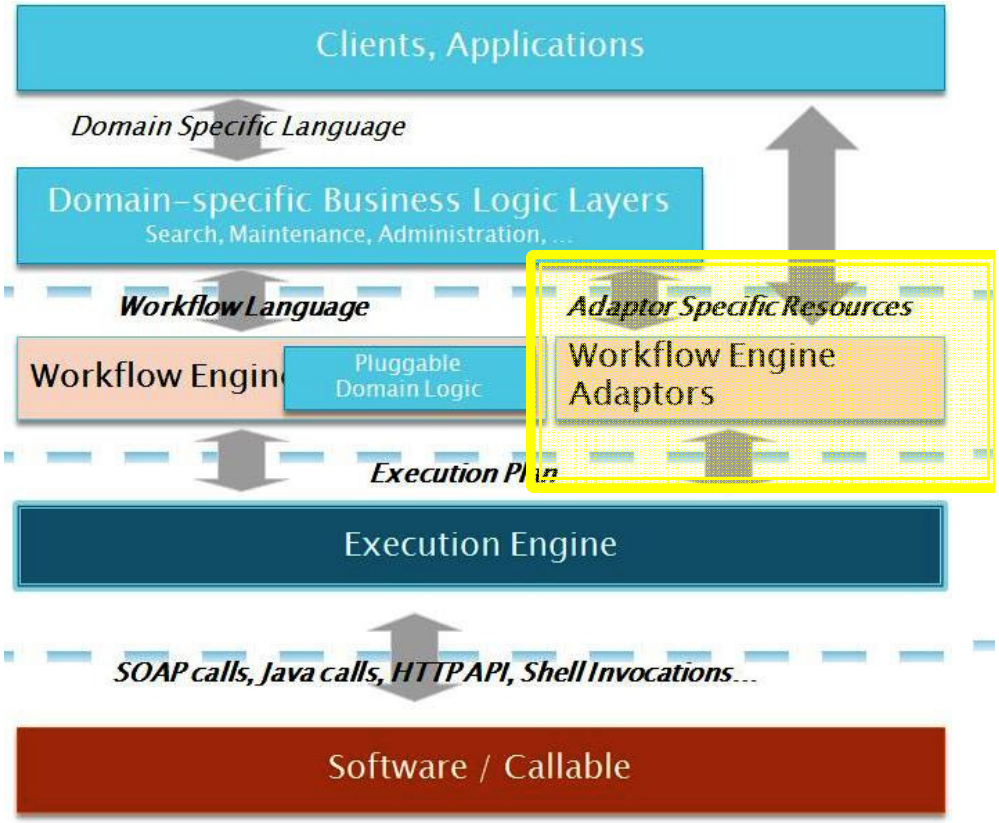


Interoperability Approaches: Process Execution [1/4]

- Process interoperability solution relies on the gCube Process Execution Engine (PE2ng) suite:
  - Invocation of a wide range of logic components: SOAP and REST WebServices, Shell Scripts, Executable Binaries, POJOs, …
  - Support for several execution paradigms: batch, map-reduce, synchronous call, message-queue, …
  - Bridges key distributed computation technologies: grid (gLite and Globus), Condor, Hadoop, …
  - Control and monitor the execution of a processing flow
  - Staging of data among different storage providers
  - Streaming data among computation elements

Interoperability Approaches: Process Execution [2/4]

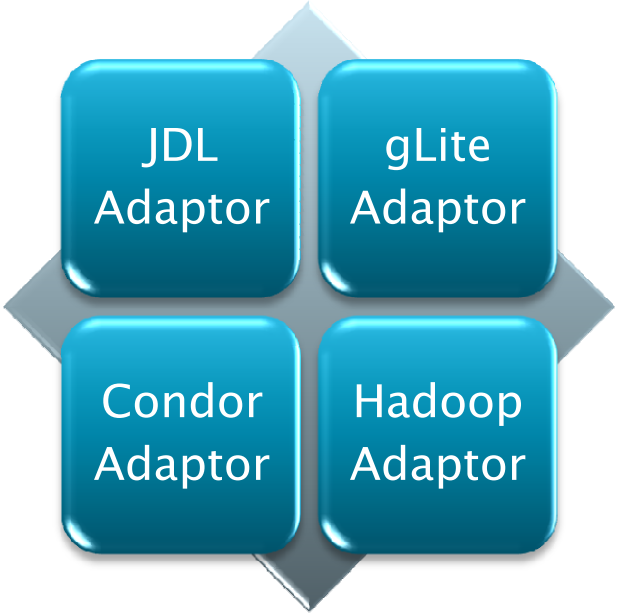
- PE2ng decouples
  - business domain and infrastructure specific logic from
  - the core "execution" functionality
- PE2ng uses Adaptors to integrate multi-infrastructure workflows



Interoperability Approaches: Process Execution [3/4]

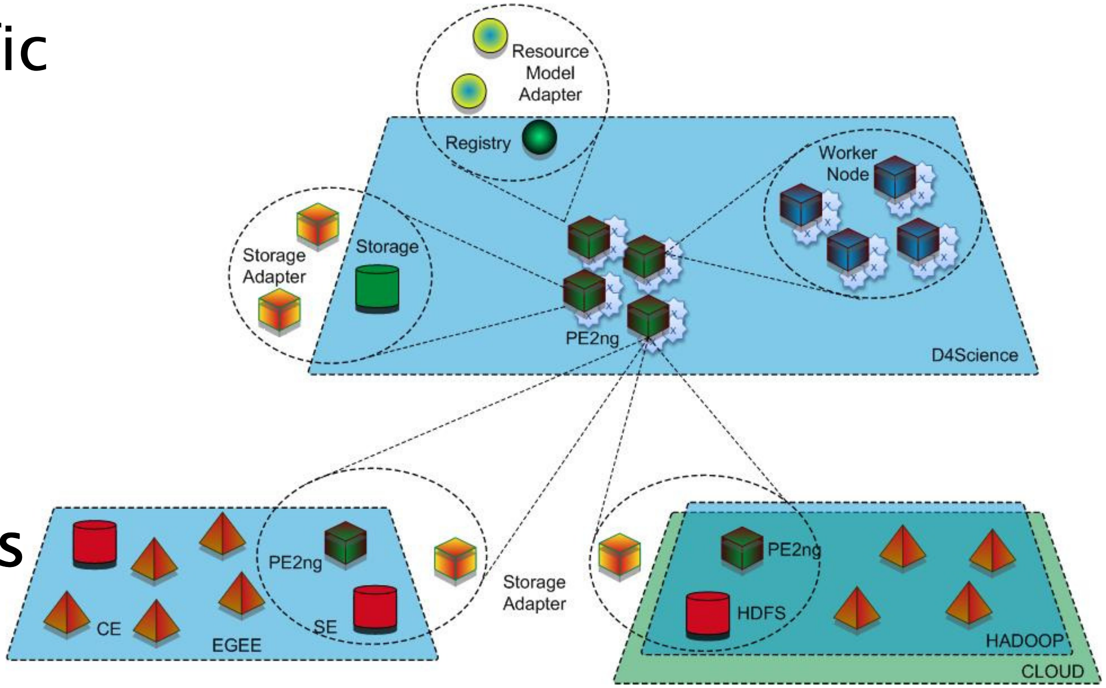
- PE2ng Adaptors parses input language and targets desired platform
  - Input contains the logic to manage execution unit profiles

  - **JDL** is the language used to specify the resources that a Grid job requires
  - **gLite** is a middleware stack for Grid computing. It is used by the largest Grid infrastructure
  - **Condor** is a software framework for coarse-grained distributed parallelization of computationally intensive tasks
  - **Hadoop** is a software framework for distributed processing of large data sets on compute clusters



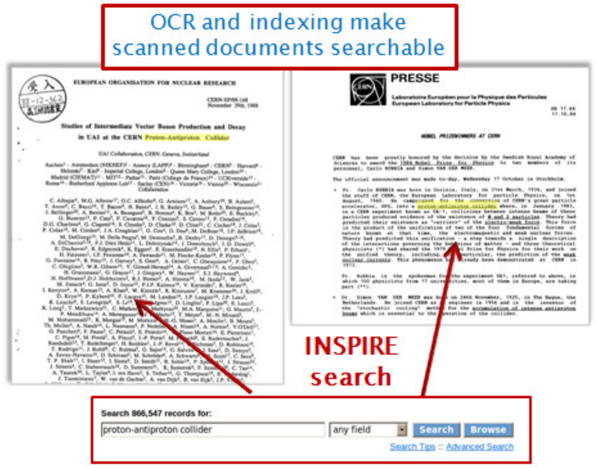
Interoperability Approaches: Process Execution [4/4]

- PE2ng Adaptors
  - operating on a specific third party language
  - translating them into native constructs,
  - allow for the creation of complex workflows that exploit several diverse technologies deployed on different infrastructures



Success Stories: Inspire

- INSPIRE is the next-generation High Energy Physics information system developed at CERN:
  - provides access to almost one million records
  - serves the worldwide HEP community of over 30,000 scientists
  - based on the Invenio digital library technology



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH

Studies of Intermediate Vector Boson Production and Decay in UA1 at the CERN Proton-Antiproton Collider

UA1 Collaboration, CERN, Geneva, Switzerland

PRESSE

Laboratoire Européen pour la Physique des Particules
European Laboratory for Particle Physics

NOBEL PRIZEWINNERS AT CERN

Search 866,547 records for: proton-antiproton collider | any field | Search | Browse

Search Tips :: Advanced Search

- Massive document OCR'ing
- Textual analysis
- Full-text indexing of large collections in a limited timeframe

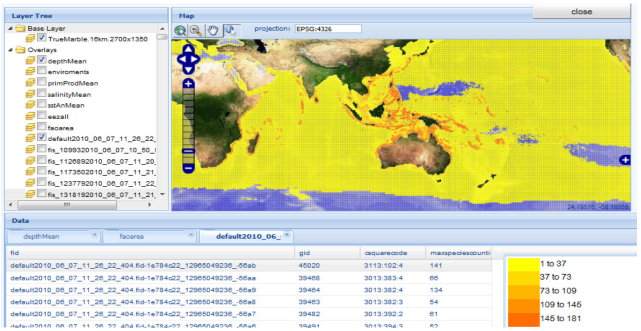
Success Stories: Inspire

- Document OCR:
  - OCR'ing jobs are executed using the PE2ng JDL Adaptor and gLite Adaptor
    - massive Optical Character Recognition (OCR) of the large HEP collections
    - textual analysis of scanned documents

- Full-text Indexing:
  - Full-text indexing jobs are executed in parallel on Hadoop clusters, interfaced through the PE2ng Hadoop Adaptor
    - Indexing of born-digital and scanned full-texts has to be re-executed every night

Success Stories: AquaMaps

- AquaMaps is an application*
  - tailored to predict global distributions of marine species initially designed for marine mammals and subsequently generalised to marine species,
  - that generates color-coded species range maps using a half-degree latitude and longitude blocks
  - by interfacing several databases and repository providers

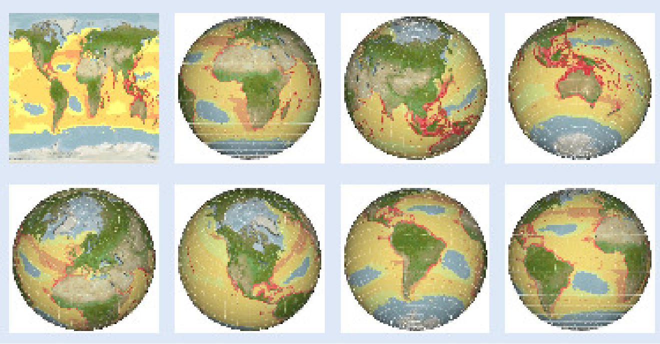


Layer Tree

- Base Layer
  - ☑ TrueMarble.16km.2700x1350
- Overlays
  - ☑ depthMean
  - □ enviroments
  - □ primProdMean
  - □ salinityMean
  - □ sstAnMean
  - □ eezall
  - □ faoarea
  - ☑ default2010_06_07_11_26_22_
  - □ fis_109932010_06_07_10_50_
  - □ fis_1126892010_06_07_11_20_
  - □ fis_1173502010_06_07_11_21_
  - □ fis_1237792010_06_07_11_22_
  - □ fis_1318192010_06_07_11_21_

fid	gid	csquarecode	maxspeciescounti
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56ab	45020	3113:102:4	141
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56aa	39468	3013:383:4	66
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56a9	39464	3013:382:4	134
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56a8	39463	3013:382:3	54
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56a7	39482	3013:392:2	61
default2010_06_07_11_26_22_404.fid-1e784c22_12965049236_-56a6	39491	3013:394:3	52

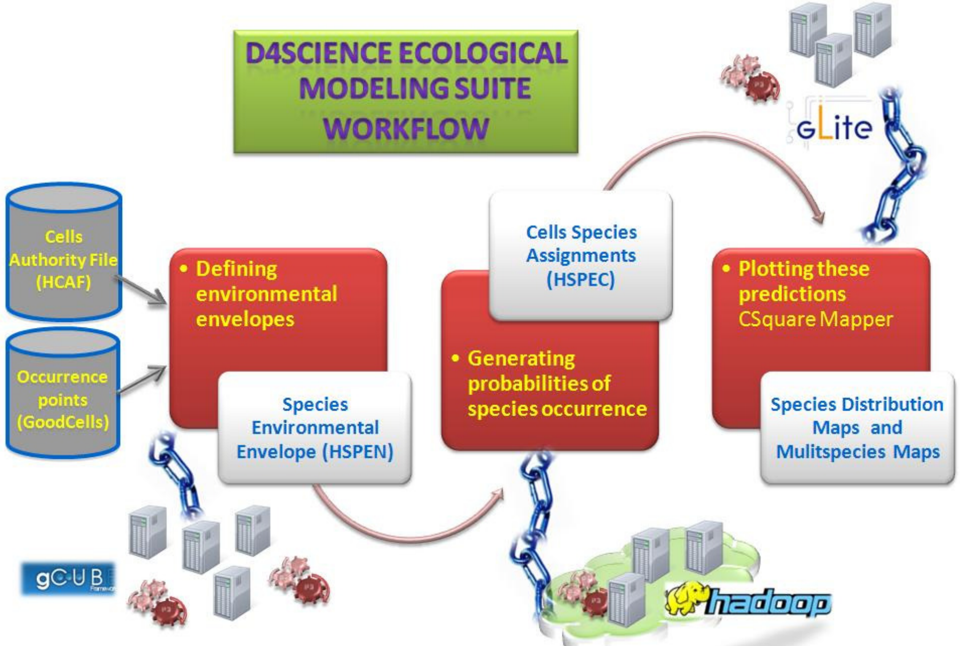
🟨	1 to 37
🟨	37 to 73
🟧	73 to 109
🟧	109 to 145
🟧	145 to 181



* Algorithm by Kashner et al. 2006

Success Stories: AquaMaps

- AquaMaps execution is based on the gCube Ecological Modelling Suite which allows the extrapolation of known species occurrences
  - to determine environmental envelopes (species tolerances)
  - to predict future distributions by matching species tolerances against local environmental conditions (e.g. climate change and sea pollution)



D4SCIENCE ECOLOGICAL MODELING SUITE WORKFLOW

Cells Authority File (HCAF)

Occurrence points (GoodCells)

- Defining environmental envelopes

Species Environmental Envelope (HSPEN)

Cells Species Assignments (HSPEC)

- Generating probabilities of species occurrence

- Plotting these predictions CSquare Mapper

Species Distribution Maps and Mulitspecies Maps

Very large volume of input and output data: HSPEC native range 56,468,301 – HSPEC suitable range 114,989,360
Very large number of computation: One multispecies map computed on 6,188 half degree cells (over 170k) and 2,540 species requires 125 millions computations (Eli E. Agbayani, FishBase Project/INCOFISH WP1, WorlFish Center)

Conclusions

- Very rich services and data collections are currently maintained by a multitude of authoritative providers
- Several standards are adopted in the same domain

- Interoperability approaches are key to exploit such richness
- The gCube framework offers a variety of patterns, tools, and solutions to delivery interoperability solutions and interconnect
  - Heterogeneous digital content
  - Heterogeneous repository systems
  - Heterogeneous computation platforms

The gCube framework is open-source and available at www.gcube-system.org

Thank you for your attention

Questions?

The work reported has been partially supported by the D4Science project (FP7 of the European Commission, INFRA-2007-1.2.2, Contract No. 212488), the D4Science-II project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and the DL.org Coordination and Support Action (FP7 of the European Commission, ICT-2007.4.3, Contract No. 231551).