Polish Roadmap toward
Domain-Specific Infrastructure
for Supporting Computational Science
in European Research Area

Reproducible e-science with GridSpace2 platform

Eryk Ciepiela, Bartosz Wilk, Daniel Harężlak, Marek Kasztelnik, Maciej Pawlik, Jan Meizner, Marian Bubak

Academic Computer Center CYFRONET
AGH University of Science and Technology

EGI Technical Forum, Madrid, 17 Sep 2013

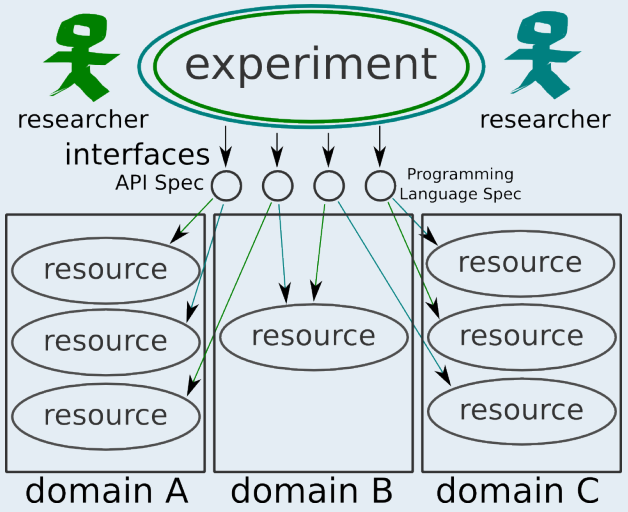
Outline

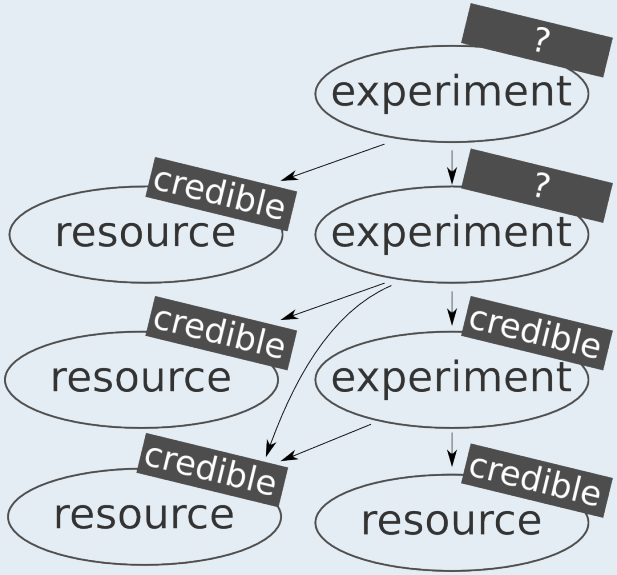
- Reproducibility problem
- Methodology proposed
- Technology proposed (GridSpace2)
- Case study (Collage)
- Challanges and future prospects
- Conclusions

Reproducibility
Motivation

- **152 entries** for **retracted** scientific papers due to **reproducibility issues** since August 2010 reported by Retracting Watch (https://retractionwatch.wordpress.com)
- **10-100k** of articles published in 1950-2004 **ought to be retracted** (Cokol et al. *How many scientific papers should be retracted?*, EMBO Reports, 2007)
  - *Nature* – retracted 30 out of 45–67 articles that should have been retracted in 1999-2004
  - Lower IF journals – lower rate by the orders of magnitude
- Publishers share concern of **credibility crisis**
  - *Article of the Future* ubmrella project (Elsevier)
  - Executable papers: publish reproducible experiments along with articles
  - *Executable Paper Grand Challange* (2011, Elsevier, won by our solution: *Collage Authoring Environment*)
- Programmers critical about scientific computing
- Science as an open enterprise (Royal Society)
- R-dimension of e-research (David de Roure)

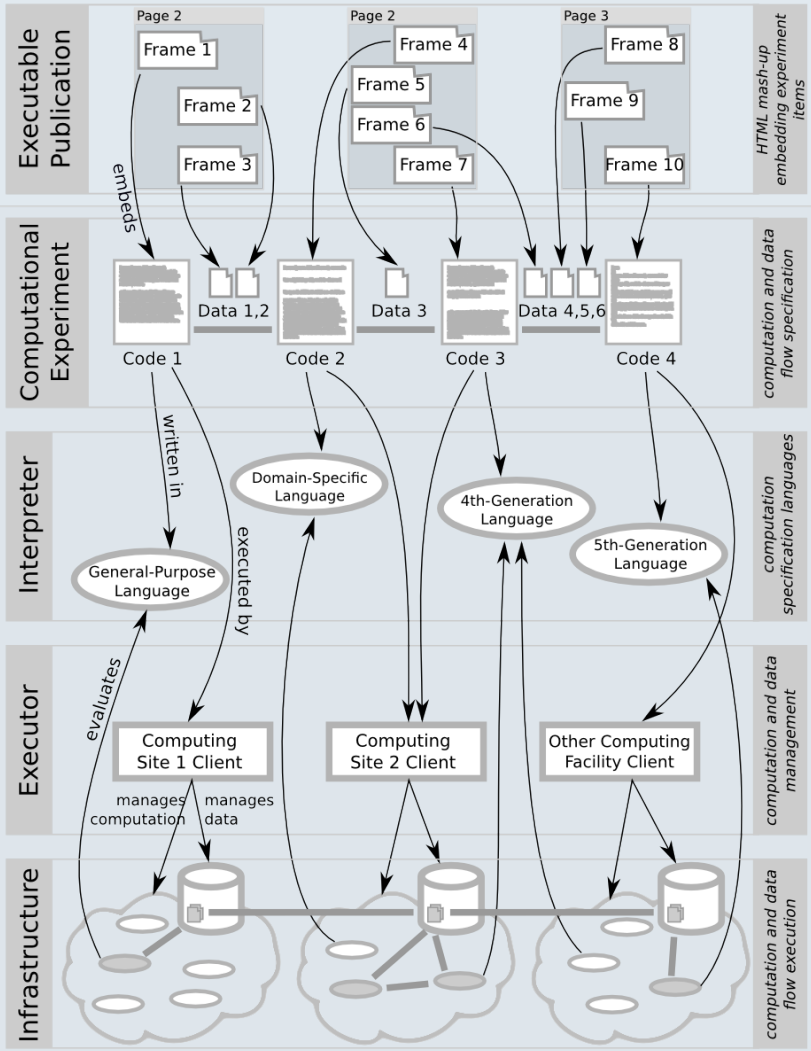
Reproducibility
Problem Statement





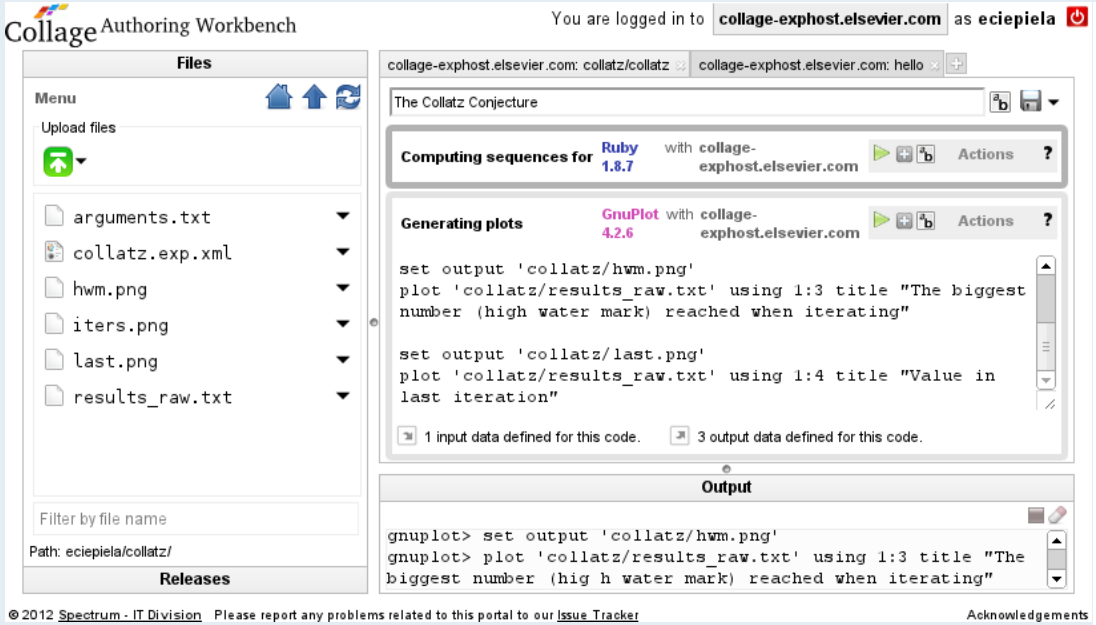
- *Reproducibility – the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else working independently. It is one of the main principles of the scientific method (…)* [Wikipedia]
- *Reproducibility – the ability to carry out the same experiment or study the other time and using the same or other resources within given time and cost constraints*
- *...other time...*
  - preservable
  - durable
  - susceptible to decay in ever evolving environment
- *...using the same other resources...*
  - coupled through interfaces
  - based on standards/specifications
  - accross administrative domains
- *...within given time and cost constaints*
  - the more straightforward the better
- *Credibile – proved, considered or evaluated as correct or derived from the other credible*
  - reproducibility gives a way to evaluate correctness thus bring credibility
  - re-use of reproducible experiments and credible resources efficient also in terms of credibility evaluation: credibility chain
- **How to bring reproducibility to e-science experiments?**

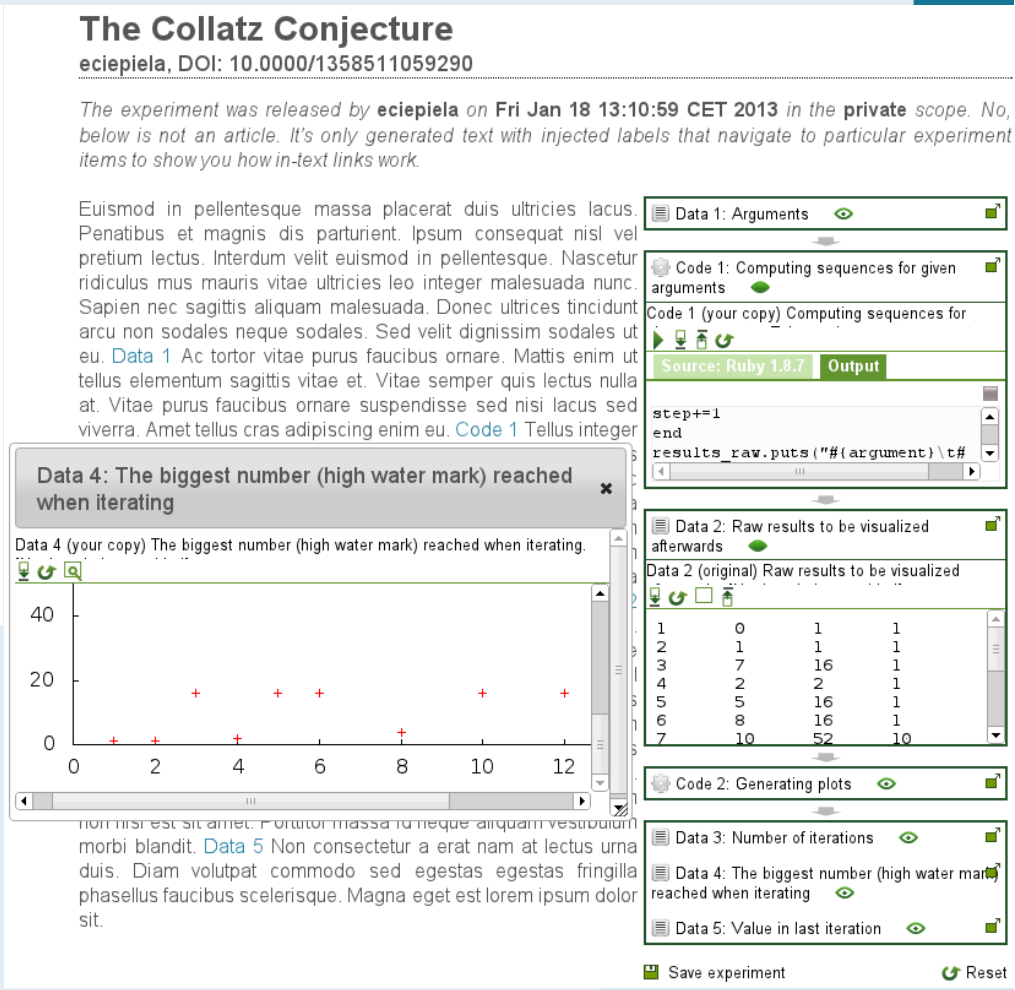
GridSpace2 Model
Methodology



- Experiments are worflows consiting of **code**, **data** and **requisite** items
- Experiments and their items are web-enabled: URL-accessible, embeddable on web sites e.g. on-line publication
- Code items are written in general-purpose, domain-specific, 4th or 5th generation programming languages that we call i**nterpreters**
- Data items are simply files or directories
- For executing code and storing data items the underlying e-infrastructures are used that we refer to as **executors**
- Experiments use e-infrastructures though abstract Executor API that constitutes an interoperability layer between various types and instances of e-infrastructures
- Experiments use interpreters without being bound to any specific interpreter implementation
- Concrete executors and interpreters' implementations are of user's choice
- GridSpace2 – web-oriented distributed computing plaftorm
- Collage – GridSpace2 + executable publications support

GridSpace2 Platform
Technology





Experiment authoring and running through a dedicated web application: **Experiment Workbench** (up)

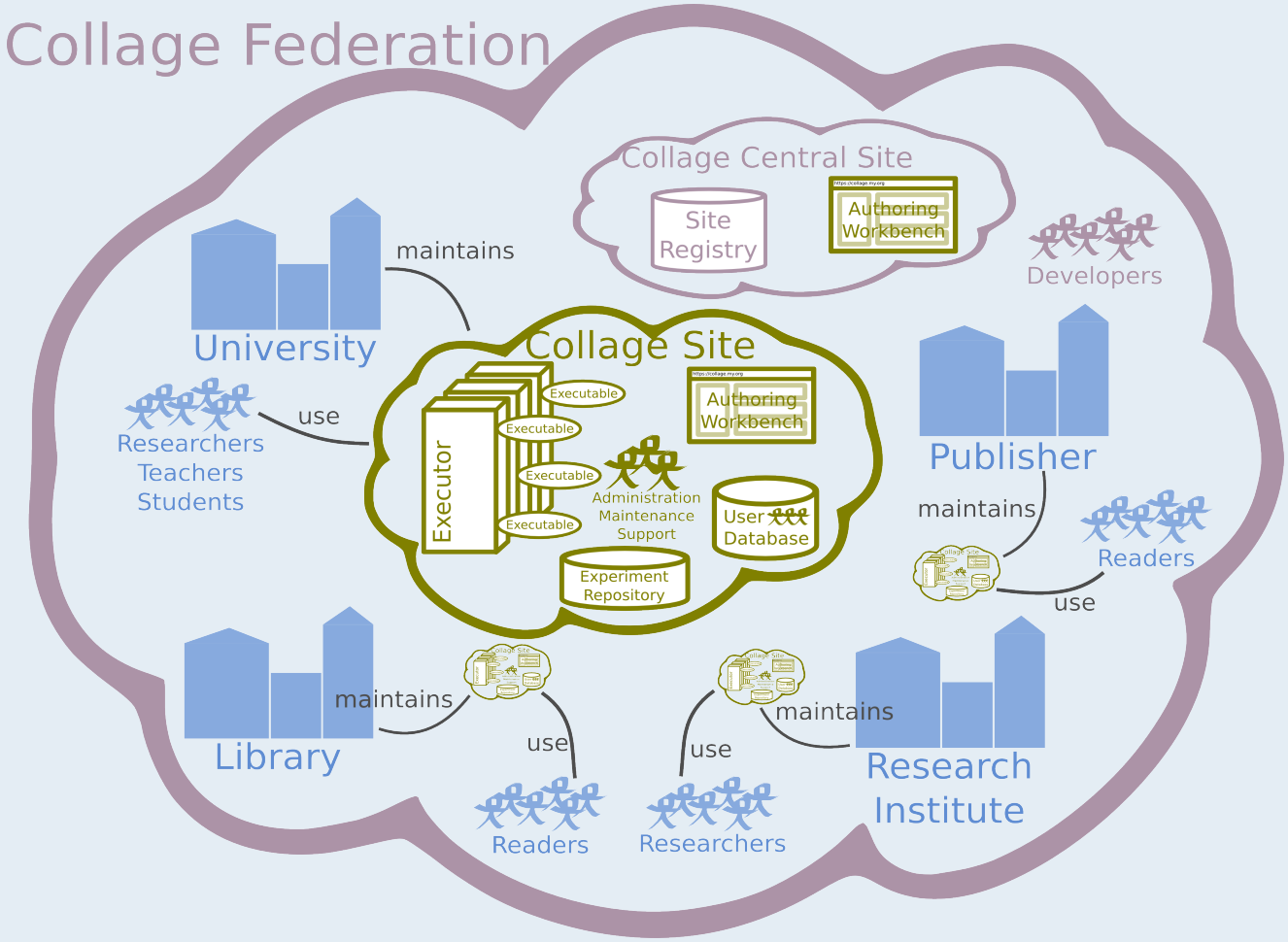
Experiment publishing on arbitrary web site, e.g. Elsevier ScienceDirect portal (right)

Collage Executable Papers
Case Study

Collage Authoring Environment for Executable Publications

- GridSpace2-based Collage Authoring Environment won Executable Paper Grand Challange in 2011
- Authoring, reviewing, publishing and executing of experiments offered by Collage proved suitable for scientific publishing in its idea
- Collage was integrated with Elsevier ScienceDirect portal so papers can be linked and presented with corresponding computational experiments
  https://collage.elsevier.com
- Special Issue of *Computers & Graphics* journal featuring 7 executable papers was released in May 2013
  http://www.journals.elsevier.com/computers-and-graphics/news/special-issue-on-3d-object-retrieval/
- (video)

Federated Collage
Challenges and Future Prospects



- Collage Site
  - University, Library, Research Institute, Computer Center etc.
  - User database
  - Experiment repository
  - Executors
  - Maintenance, support
- Central Collage Site
  - Registry of interpreters and federated Collage Sites
  - Experiments portable between Collage Sites
- “Publish globally – compute locally” model
- Collage/GridSpace2 software as public open-source project
  - Further development voluntary
- Collage/GridSpace2 software shipped to Collage Sites as virtual machine appliance

Conclusions

- Reproducibility is an inherent part of scientific method
- E-science suffers from lack of reproducibility what causes credibility crisis in scholarly papers
- Poor quality of scientific programs leads to retracted papers so they also must be subject to review
- Credibility of e-science experiments must be efficiently evaluated
- New methodology is needed to embrace above mentioned concerns – GridSpace2 model is the example
- New tools are needed to aid such methodology – like GridSpace2 platform
- GridSpace2/Collage platform is scalable and their instances can be federated
- With federated instances of GridSpace2/Collage platform we can achieve reproducibility accross administrative domains and scientific communities

GridSpace2