Candidate technologies for review

This page contains a list of the items being considered and reviewed as part of the ENVRIplus technology review. The entries may refer to general *are as* that need to be covered, specific *technologies*, specific examples and *implementations* of those technologies, or specific examples of the *applicatior* of those implementations. For example, the general *area* might be *data-intensive workflow systems*, the specific technology might be *data-streaming workflows*, the implementation might be *dispel4py*, and the application might be its use *correlating traces from 500 seismometers on an hourly basis.*

Entries should be made on this page when

1. A member of the ENVRIplus notes that a topic needs to be covered, or
2. A member of the ENVRIplus project commits to, is in progress of, or has reported a review of the topic.

The purpose of these entries is to keep each other informed, and to reduce the risk of unnecessary duplication. It will also help track progress.

Entries should be logically grouped, for example under the six pillars of theme 2, or the three cross-cutting areas of investigation, namely:

1. Data identification and citation
2. Data curation including quality control
3. Data, Method, Process and Resource Cataloguing
4. Data Processing including transformations for integration
5. Optimisation of data-handling to reduce human effort or reduce resource use
6. Provenance including tools and mechanisms exploiting provenance
7. Community support including Virtual Organisations and CSCW
8. Architecture including mechanisms to meet non-functional requirements
9. Linking model
10. Reference model including extensions to ontologies to improve description
11. Other topics

Within that categorical structure related issues, such as the drilling into more detail on a topic, should be grouped.

Each entry should contain the following information:

1. A title by which this topic will be known (Name)
2. A URL where treatment of this topic can be found (when it is available) i.e, link to another page in this wiki
3. The person(s) making the entry
4. The date the entry was initially made
5. A short description of the topic, delimiting its scope
6. Related topics
7. The current status: (a) requires investigation, (b) will be investigated, (c) is being investigated, or (d) has been reported.
8. For b, c or d above the person(s), group, ENVRIplus task or WP that is undertaking the work.
9. Its category, from the list above, or a new category name, which should be gathered under *other topics.*
10. Its depth, from *area, technology, implementation or application.*
11. Its form, e.g. *Protocol*, *Distributed framework*, *SW system, SW library*, *HCI framework*, *Tool*, *metadata framework*, *ontology*, *standard*, *authori tative report*, ...
12. Where relevant its maturity, e.g. TRL for software
13. Its potential source
14. Why this topic is critical and needs attention

The topics follow, with Category headings dividing the list.

(copy and paste the text below and then complete the entry. Leave the text below as it is for someone else to copy and use)

>>insert topic here with Heading 3 style<<

1. A URL where treatment of this topic can be found (when it is available) i.e, link to another page in this wiki
2. The person(s) making the entry
3. The date the entry was initially made
4. A short description of the topic, delimiting its scope
5. Related topics
6. The current status: (a) requires investigation, (b) will be investigated, (c) is being investigated, or (d) has been reported.
7. For b, c or d above the person(s), group, ENVRIplus task or WP that is undertaking the work.
8. Its category, from the list above, or a new category name, which should be gathered under *other topics.*
9. Its depth, from *area, technology, implementation or application.*
10. Its form, e.g. *Protocol*, *Distributed framework*, *SW system, SW library*, *HCI framework*, *Tool*, *metadata framework*, *ontology*, *standard*, *authori tative report*, ...
11. Where relevant its maturity, e.g. TRL for software
12. Its potential source
13. Why this topic is critical and needs attention

Data identification and citation

Data curation

Cataloguing

Data processing

Scientific Workflow Management Systems:

- Meandre;
- Kepler;
- Pegasus;
- Swift;
- Taverna;
- ...

gCube Data Analytics Platform;

Optimisation

Provenance

PROV

1. A URL where treatment of this topic can be found (when it is available) i.e, link to another page in this wiki - to be made
2. Barbara Magagna
3. 20160118
4. An overview of the Prov Family of Documents and its relevance for Provenance
5. A model for Provenance
6. The current status: (b) will be investigated
7. Barbara Magagna, Thomas Loubrieu, Keith Jeffery and whoever wants
8. Provenance, underlying model
9. Implementation
10. *ontology*, *standard*
11. Where relevant its maturity, e.g. TRL for software
12. www.w3.org/TR/2013/NOTE-prov-overview-20130430
13. This is one of the most cited ontologies in this areas and should if not used at least be a basis for comparison

CERIF

1. A URL where treatment of this topic can be found (when it is available) i.e, link to another page in this wiki - to be made
2. Barbara Magagna
3. 20160118
4. An overview of the CERIF model and its relevance for Provenance
5. A potential model for Provenance
6. The current status: (b) will be investigated
7. Barbara Magagna, Thomas Loubrieu, Keith Jeffery and whoever wants
8. Provenance, underlying model
9. Implementation
10. *concept model*, *standard*
11. Where relevant its maturity, e.g. TRL for software
12. http://eurocris.org/cerif/main-features-cerif
13. This is one of the most cited concept models in this areas and should if not used at least be a basis for comparison

Community support

Architecture

Data-Intensive Federations

1. A URL where treatment of this topic can be found (when it is available) i.e, link to another page in this wiki
2. Malcolm Atkinson
3. 18 January 2016
4. Data-Intensive federations (DIF) are formed to enable practitioners to have easier access to *dynamic* and *evolving* data that is *owned* and *pro vided* by *multiple independent organisations* some of whom may be *partners* in the DIF. A DIF needs to be long-lived to enable its many users to depend on its services. During its lifetime the provisions, priorities, data organisation and services of the data providers will evolve, as will the requirements and activities of its user community. The target of gaining benefit from the improvements in data acquisition, data preparation and data curation happening contemporaneously in provider organisations and the requirement to handle dynamic data so that uses can have response horizons vary from almost immediate to very long term, differentiat DIF from Digital Asset Management (DAM), which helps practitioners develop static collections of under their own control. An effective DIF delivers a holistic and comprehensible view of the relevant data to its users, it facilitates the specification and application of dynamic data integration strategies and it permits effective working with a wide variety of data analysis systems, problem-solving and development tools and all of the functional and non-functional aspects of a DAM. In particular it supports a *Virtual Research Environment* (VRE), which is underpinned by a *Virtual Organisation* (VO) that administers identity, membership of groups, allocation of roles, and hence of authority to use data and resources. The implementation needs

to implement these rules with proper security and accounting across the provider and partner organisations. The DIF will offer computer-supported collaborative working (CSCW), e.g. sharing workspaces, collaborating on developing scientific methods, data handling processes, agreed data organisations, vocabularies, ontologies, user-driven metadata, and so on, with specified scope. However, these will extend to dynamic access, handling and integration processes in the DIF context that are designed to be reused repeatedly on demand. To achieve sustainability, most of this work should be constructed at an abstract level not bound to underpinning technologies, platforms and computational resource provisions. The *scientific gateway* should *dynamically map* user actions (often performed using tools that call the science gateway's API) onto executable data-intensive workflows or distributed queries that are deployed across the distributed infrastructure to deliver the required result. With sufficiently high-quality descriptions of the platforms, components, services and data these mappings can be largely automated. It is necessary to develop a good architecture for DIF, that is reusable for many DIF. Otherwise DIF (and even DAM) will become unsustainable as their context evolves and as the number of pairwise interactions between components grows at an order N^2 rate.

5. Related topics: Digital Asset Management (DAM), Data Integration, Virtual Research Environment (VRE), Computer-Supported Collaborative Working (CSCW), Virtual Organisations (VO), Virtual laboratories, Science Gateway, Dynamic Mappings, Data-Intensive Platforms, Data-Intensive Workflows. Distributed Databases and Queries.
6. The current status: (a) requires investigation, (b) **will be investigated**, (c) is being investigated, or (d) has been reported.
7. For b, c or d above the person(s), group, ENVRIplus task or WP that is undertaking the work. **Malcolm Atkinson, Alex Hardisty and Keith Jeffery**
8. Its category ***Architecture***.
9. Its depth, from ***area***, *technology, implementation or application.*
10. Its form, e.g. *Iterative development of an architectural style and at least one candidate implementation*
11. Where relevant its maturity, e.g. TRL for software: *Novel proposal needs attention*
12. Its potential source: *being investigated*
13. Why this topic is critical and needs attention: Without it, work on e-Infrastructure to support RIs and on tools to aid all of their practitioner roles will run into a complexity barrier and become unsustainable. This is particularly the case where some of the data used is collected and used for other social, political or commercial purposes, and the RI needs to sustain effective working with provider organisations that have other priorities.

Linking model

Reference model

Other topics