

2007 Montpellier Platform Development Workshop

Overview

A GCP platform developer's workshop will be held at Bioversity International in Montpellier, France from 27-30 November 2007. The primary objectives of the workshop is to review progress in the implementation of GCP platform standards to GCP target use cases, to provide tutorials on GCP standards to developers less familiar with these standards, and to undertake some additional planning for 2008 GCP platform development.

Meeting Notes

Monday 26 November

Present: Mr Max Ruas, Mr Milko Skofic, Mr Mathieu Rouard, Dr Manuel Ruiz, Dr Georgios Pappas, Mr Marcos Costa, Mr. Subramanyam Goli, Dr Richard Bruskiwich, Ms Mylah Anacleto, ()Brigitte Courtois, ()Nicolas Roux, ()Elizabeth Arnaud, Claire Billot

AM - Review of SP1 Use Case - Progress from 2007 & 2008 Use Cases

Manuel Ruiz presented a progress report for 2007 GCP platform SP1 use case work

Reference: file- ATRSP4Task2007-16.doc

General Action Items:

- Brigitte: proposal to send request to geneticists to provide data. difficult to test tools without data inside.
 - For IRRI, need ask Dr. McNally about the status of current data on Rice.

Manuel presented 2008 Workplan

Reference: file - GCP Information Platform 2008.doc

- Visualization of data: use of Data Transformer to wrap Structure, R and Instruct on HPC -- ask Martin for update regarding this.
- Develop more GCP data sources for passport dbs developed by Bioversity
- One of the issues that Richard, Milko, Martin and Matthieu will try to address this week is the framework involving MOBY, TAPIR and GCP data sources
- emphasis in 2008: passport data seamlessly integrated with genotyping data in the platform
- ask Thomas to enter in wiki an update about the data quality project (as a follow-up of what was agreed with during the data quality workshop in 2007) action needed
 - Genotyping Data Quality Workshop, 2007, Los Banos
 - Passport Data Quality Workshop, 2007, Rome
 - Technical manual on passport data quality

Update from Claire:

- for sorghum -- start in January 2008
- Richard: how to compare two data sets considering inconsistencies are expected
- Claire: tools should be designed, (available c/o Manuel multiple experiment results for a combination of marker but works only on one platform)
 - export to IRRI ok because IRRI is using the same platform as CIRAD(e.g. SSR markers, EBI system)
 - meeting last year in IRRI, asked Guy Davenport to update templates to have internal measures to ensure quality of data
 - no template for DArT and SNP data templates - ask Guy about this action needed
 - CIMMYT has same need
- Brigitte: follow-up Australian team working on data templates (probably also ask Guy) action needed
- Claire: data to be stored using data templates

- dealing with genetic data and passport data
- work on these data: data quality - how can we be confident about the data
- genotyping point of view and passport point of view
- it was discussed in netherlands (thomas is best source), data quality concern is top priority
- integrate it in the way we are working
- concern on analysis: genetic diversity point of view
- need to have a connecting tool between the template and the GCP repository and informatics tool we are using
- some tools do not exist, e.g. Darwin developed because more features needed
- genetic analysis
- Richard: for data quality: algorithm - how geneticist can give assurance about the quality of data, is it based on judgment (user annotation using ontology?) or is it something computable (e.g. like a sequence Phred score, based on a computer model of quality?)
- Claire: meeting didn't go deeper (assessment of quality).. to be done under project c/o Thomas
 - data quality problem: run twice and different results are obtained; Congruency (identity of results) check on subsets of data
- Richard: not automation of data quality checking but give end-user facility to...
 - quality indicators, annotation: consistent annotation that suggests a dictionary of well-understood ...
- What is expected at the end of 2008:
- Claire:
 - have real data, access to raw data (e.g. gene diversity: query filter according to different fields, searchable data, template data fields, type of machine experiment was done, etc.)
 - e.g. samples run in 2006 and 2005.. if more confident about 2006, use only 2006 data. Compare both data. If congruent (95%), infer 2005 data are okay, consider even only run once.

- Richard: what percentage of samples have discrepancies... sounds like LIMS: need to know about experimental batches (i.e. sample plates, gel runs, etc.) We have not mentioned LIMS in the use case. Does the GCP template record such information (e.g. experiment id as a gel run or a single run on an EBI system) -- a batch number?)
- Claire: in templates, some fields are mandatory, some are not. Experiment (batch) number is optional
- Richard: appears that the fundamental measures of quality is congruency between data batches.

Summary:

- for SP1, data quality emphasized
- further elaboration of what is expected
- Claire: tools for analysis also needed

--- Break for lunch ---

PM - Review of SP2 & SP3 Use Cases

Richard Bruskiwich presented a report on 2007 progress on the GCP platform SP2 use case

Refers to files: file- ATRSP4Task2007-16.doc and GCP Information Platform 2008.doc

- for SP2, the second concern is the client application (Koios)
- discusses components - (1)for microarray data, (2)web portal, (3) viewers
- Marcos Costa was involved in Apollo, IRRI-hosted Canadian summer student James Wagner for Cytoscape, -- basically connect these tools to Koios
- Invoke 3rd party tools from the web using Java Web Start - Andrew Farmer from NCGR will give us further advice on how to do that in this meeting
- some tasks to spill over to 2008
- some web elements in dayhoff to be used in Koios

- Cytoscape is a Java network data visualization tool, can be easily integrated, maybe generic enough to do the job for visualizing other non-gene clusters of data (i.e. accessions)

Reactions to presentation: Brigitte: looks like a lot of work

Response: Richard: after a slow start with staff recruitment, considerable momentum from 2007, most of the key implementation issues resolved; IRRR team now organized into two teams, "back end" (GCP DataSource developers) and "Front End" (GCP web/3rd party tool developers).

Marcos

- working on Apollo
- Need to further clarify how GCP data type attributes mapped to data

Mylah

- no need to lay out all requirements at once, can be gradual, do in iterations

Richard

- rapid iterations promised at start of project, but only now fully feasible given that formal GCP domain model and platform framework, and staffing, are in place.

On behalf of Guy Davenport, Richard Bruskiwich briefly presented notes about GCP platform SP3 use case

- some discussion on iMAS
- for 2008, the CIMMYT-led SP3 use case and ICRISAT's concept on what it wants to do: can these somehow converge into one project
- with ICIS it is also possible to query pedigrees
- in 2007, on the CIMMYT-led SP3 use case, no detailed design & implementation was undertaken. in 2008, the challenge is to implement the prototype (with ICRISAT)
- no duplication of effort - suggest development of a Java (Swing?) data consumer that can be used in different places (In Genomedium? In Koios?)
- discussion on deliverables c/o Guy, Graham and Jayashree

Tuesday 27 November

Present: Mr Max Ruas, Mr Milko Skofic, Mr Mathieu Rouard, Dr Manuel Ruiz, Dr Georgios Pappas?, Mr Marcos Costa?, Mr. Subramanyam Goli, Dr Richard Bruskiwich, Ms Mylah Anacleto, Andrew Farmer, Pierre Larmande Martin Senger

Meeting Notes

Discussion of Objectives (see above agenda)

- Objectives and expectations of the meeting:
 - Review of GCP platform collaboration - available resources and strategies
 - Generally: that all developers in attendance become fully informed about GCP platform - DataSource, DataConsumer, DataTransformer - API, DataSource validation, and about current implementation frameworks (shared libraries) currently available
 - Tapir/BioMoby - review and agreement on GCP-compliant BioMOBY data types and services required; clarification on technical strategy and DataSource strategies
 - ICRISAT: want to wrap ICRIS as a GCP DataSource - knowledge of GCP domain model (data types), and framework to be used
 - 3rd party software integration (with JavaWeb Start)
 - CIRAD DataConsumer and DataTransformer development: specific GCP domain model and use case usages
 - EMBRAPA/Genoma: specific GCP domain model and use case usages
 - Integration of GCP platform (web) application integration - CIRAD web interfaces, Koios and iMAS.

Review Discussion about GCP DataSource API

- Do DataTypeAttributes need a data "type" (getDataType():String)
- Should DataSource.metadata return test data for PantheonValidator test

data

- BioMOBY NameSpaces:
 - Use LSID for GCP domain model objects (as prescribed in Pantheon docs Identification)
 - Split LSID in two parts for GCP MOBY objects i.e.

`<GCP_SimpleIdentifier ns="irri.org:IRIS.Germplasm" id="110:1"> ... </GCP_SimpleIdentifier>`

has a corresponding LSID:

`urn:lsid:irri.org:IRIS.Germplasm:110:1`

Action Items?

- Updating of DataSource "best practices" is required:
 - ALL_REASONABLE searchable attribute (R. Bruskiwich)
 - GCP BioMoby <=> LSID interconversion (R. Bruskiwich)
 - Prescribe that DataSource default (empty) constructors not be intensely resource consuming (M. Senger)
 - Update ICIS and other IRRI datasources to follow this best practice (M. Anacleto => IRRI team)
 - Harmonization of GCP DataSource find() use case documentation(?)
- Web interface to Validator (M. Senger)
- MOBY Environment: change directory structure to partition services by "web service provider" (M. Senger)
- Need for Ontology web services (to be elaborated further by Milko and Richard)
- Should the DataTypeAttribute interface include a method to return the attribute datatype (getDataType():String)?

Other Miscellaneous Facts

- Taverna has new capabilities to be exploited by the GCP: see myexperiment.org
- SoapLab2 is now ready for release; nice new easy Ant installation and

HTML web form generation framework available; interoperable with Taverna

Wednesday 28 November

Present: Mr Max Ruas, Mr Milko Skofic, Mr Mathieu Rouard, Dr Manuel Ruiz, Dr Georgios Pappas, Mr Marcos Costa, Mr. Subramanyam Goli, Dr Richard Bruskiwich, Ms Mylah Anacleto, Andrew Farmer, Pierre Larmande, Martin Senger

AM - Inventory of available data source

Richard discussed the list of available data sources from IRR1
reference: <http://pantheon.generationcp.org/components/datasources/index.html>

Other (important) topics

Manuel: it is important that the pantheon docs pages are up-to-date
Topic: Pantheon documentation update

- outdated contents identified (many more, just examples):
 - releases
 - inventory of data sources: remove ds that are not working
- how to make update manageable
- Martin: front page should still contain the main components
- use a content mgt system: resources available to make this possible?

Martin: next action items:

- create prototype of pantheon documentation in joomla!
- test it, agree on it

Martin: what would be the policies for content management?

Matthieu: (about login accounts) proposes that existing accounts in cropwiki be created in joomla! version of pantheon docs?

Moby registry

- We have been having frustrations connecting to cropwiki server in IRR1

Martin will document decision on a new way for datasource discovery mechanism (list of jar files)

PM

Synopsis about using available IRR1 GCP DataSources

List of IRR1 DataSources

DataSource	Current Maven groupId/artifactId/version	DS or DS F(*)	Class Name in SPI File	Configuration Particulars
Chado (General) Client	org.gmod.chado/chadods-client/1.1.0	DS	org.gmod.chado.datasource.client.GCPChadoDsClient	n/a
Chado Ontology	org.gmod.chado/ontologyds-service-client/1.5.0	DS	org.gmod.chado.cv.datasource.service.client.GCPChadoOntologyDsServiceClient	n/a
ICIS (e.g. IRIS, IWIS, IMIS)	org.generationcp.osiris.springdatasource/springds-client/1.3.6	DS F	org.generationcp.springdatasource.client.factory.HttpInvokerDataSourceFactory	SpringDSClientConfig.properties should be located on the Java classpath and include lines with names of the configuration file for various ICIS installations (e.g. IRISClientConfig.xml for IRIS, IWISClientConfig.xml for IWIS, IMISClientConfig.xml for IMIS) which should also be on the Java classpath
BioMoby Client	org.generationcp.moby/GCP-Moby-Client/1.2.4	DS F	org.generationcp.moby.client.MobyClientDataSourceFactory	MobyClientDataSource.xml + all web service provider xml files (e.g. GREENPHYLMobyProvider.xml) should be located on the Java classpath
BioJava	org.generationcp.osiris	DS	org.gmod.e	n/a

	s.biojava/ embls-client/ 1.0.3		mbl.datasou rce.client.E MBLDsClien t	
GDPC	org.generationcp.osiri s/ gdplib/ 0.1.8	DS	org.generati oncp.gcpgdp c.SpringDat aSource	n/a
MAXD	org.generationcp.ma xd.datasource/ /maxd- datasourceClient/ 0.0.3	DS	org.generati oncp.maxd.d atasource.M axdDataSou rceClient	n/a

(*) DS = DataSource class name in DataSource SPI file, DSF = DataSourceFactory class name in DataSourceFactory SPI file

[edit] MOBY Client Data Source

Who will benefit from this? who are going to try it out: Milko, Mathieu, Georgios, Manuel, ... almost all

Has this been tried? -- before the ARM, we were able to discover moby client data sources

Mathieu: it is important that we have a stable domain model so conversion library would also be stable and not difficult to maintain

Richard: model is end-driven. since march, only incremental changes driven by the development.. only when used do we discover there are changes that need to be done

Martin: in moby registry, every data type has version but registry does not actually keep all of them when service uses data type...

Richard: we have to deal with the possibility that there are two versions of feature for example.. the conversion library will have to specify which version actually to use inter-conversion will have to worry about that

Martin: i do not think so in java, we deal with java objects.. maven handles the versioning of libraries explore possibility of opening it up with mark wilkinson to ask the modifications in the registry. although using lsids, registry does not keep older versions

version automatically incremented

convert non-gcpmoby objects to GCP objects (e.g. output from non-gcp services to GCP objects) yes... must be possible

Action for Martin: consider the possibility of including version checking into BaseClient (MOSeS)

4:15 PM Maven as tool for development

Martin discussed how maven is used in GCP component development

Action from everyone who use configuration files: configuration files that can be modified should be available in the jar file as a file with name ending in ".template"

jar files used in moby development:

- biomoby data types -- already in the maven repository
- put it differently for canadian and irri registry
- For mylah(todo 11/29/2007): initiate creation of gcp repository group :
 - use org/generationcp/moby/irri/
 - use org/generationcp/moby/canada/
- put biomoby-datatypes.jar

Action from Martin: administer Archiva (accessible via <http://maven.generationcp.org:8080/archiva>)

Thurs 29 November

Plan for AM - Breakout sessions

8 h - 11 h

- ICRIS database - Mylah, Subbu
- Apollo - Marcos, Richard
- Validation of CIRAD DS - Martin, Manuel, Xavier

- Tapir Moby - Milko, Mathieu, Richard

11 h - 13 h

- Moby DS Package - Mylah, Manuel, Xavier
- Java web start - Martin, Marcos, Andrew

??ISYS - Andrew, Manuel

Friday 30 November

Morning

1. Joomla! embraced as new target for management for Pantheon project documentation

2. What bug tracker: CropForge versus Jira bug trackers... Strong opinion that CropForge should be

embrace.

3. should GCP platform developers use GCP comm or Pantheon bugtrackers, et al.

4. Training videos good for end users, NOT developers, unless a PowerPoint presentation

introduction about a technology (bird's eye view, presentation... introduction to talk, perhaps

for wider, general audiences.

If Thomas has resources to develop special documentation support fortutorial development for SP4

products, then perhaps this will be a good line to explore.

5. Best practices for SP4 software releases needed soon (but not yet?). Packaging and installation

of SP4 products is an issue to be tackled soon.

- Web applications? - not too frequent
- standalone applications? - Genomedium, support???

- What kind of installers do you need?

InstallAnywhere? Open source installers? "NSis"
<http://nsis.sourceforge.net/Download>

6. What is not going well enough in our GCP platform project (good, bad and the ugly)? How can we

do things better?

- We need face-to-face GCP platform workshops at least twice a year
- we can *try* to have more videoconferencing/teleconference (multi-site) in between
 - "Webex" - third party commercial and other (Thomas??)
- Better public advertising of SP4 products and development experience/best practices; refactor

Pantheon docs into a more layered, technical documentation, better communication of project

rational

- Why should you use Pantheon?
- What is available, for what specific purpose?
- What is the status of different components?
 - JOOMLA! news? Thomas? Graham? other project PI's
- Better interaction with external collaborators to engage with external partners (NCGR, GDPC).
- We need to have more success stories about sharing GCP platform technology, without this, we are

• dead - the project has "failed". We are close but we have not yet succeed. We don't yet share

enough? Technology is not yet shareable across the GCP partner sites. GCP platforms are still silo

developments in partner sites. Some libraries are *nearly* shareable...

- How can we do this?
 - Can Strategic collaborations be established between core GCP development partners to push

•libraries, tools toward shareability: coding for generalization, improved technical documentation, troubleshooting software for shareability, stability and robustness of projects (stable production management)

- Consolidate core GCP platform libraries, better documentation of the libraries
- Performance of software should be considered
- We collectively need to have better change management, that understands impact of changes
 - e.g. versioning system for Domain model: propagate model changes into narratives, MOBY, etc. Serialization number.
 - Library changes may also be an issue (e.g. serialization)

•After noon

- MOBY DS package presented by Mylah
 - Martin suggested
 - to use spring to generate automatically some part of the configuration file for the moby client package configuration.
 - that registry endpoint is useless because information about service is already there.
 - to have a separate jar file for target data types translation from domain model (currently in ceres moby)