

SaoPaulo ARM Review

GCP Annual Research Meeting: SP4 Workshop -
September 17, 2006

Integration of the HPC facilities in the GCP Toolbox (Anthony Collins)

Rapporteur: Guy Davenport

Comments

- Hardware expansion or upgrade not possible
- Need to increase cost benefit to GCP users
- Continue with use case development
 - Take presented use cases and extend user group
 - Document
 - Test
 - Extend to other users in GCP
- Web site, summary site for project (not compete with cropforge) e.g. hpc.generationcp.org
 - By the end of the year
- CIP etc will provide support, using a duplicate HPC for test upgrading
- Move our apps to large scale systems (hired or borrowed systems)
 - Platform independence of GCP HPC systems, between (IRRI, CIP, ICRISAT) and out side of GCP (ILRI and others?)
 - Secondary goal
 - Future proofing
- Need to look at LEGES (Enabling Grid for E-Science) as an example

- Soaplab **DataSource** now available (IRRI)
- Grid of HPC we have the technology, but not the connection speed (CIP, IRRI, CIMMYT). Need I2 connectivity.
- Need user to access systems for non-HPC sites

Use case of Structure (CIP/CIMMYT)

Theo's Summary Comments

- New and expanding use cases
- Experiment with compatibility with over systems to get experience
- New project on bioinformatics resource coordinations

Domain Modelling and Ontology Development (Richard Bruskiwich)

- Rapporteur: Graham McLaren

Version 1 of the Domain Model has been released in July and changes will only be made to correct errors which prevent implementation as we develop the platform and other model dependent tasks.

Various projects (GCP sites and non GCP partners) using passport data have been using an XML schema model from the templates task. Now that version 1 of the Domain Model is released these will be upgraded to XML schema derived from the Domain Model during the rest of 2006 (Guy will derive the XML schema and Samy will coordinate its implementation in the existing applications using passport models). The derived XML schema should be published as an alternative documentation of the Domain Model.

Jayshree asked about how the DM was translated into Moby Objects. Martin responded that this is a necessary part of the documentation and if it is not clear now the documentation will be improved. (Martin will liase with Jayshree to see that this process is well understood and documented).

Sami: GPG2, world bank upgrade, 1st Jan 2007, will have big data standardization, phenotyping part , 7 crops in common, 6 relevant to GCP. 7 initiatives, crop experts, review descriptors, enrich descriptors. Other crops to come later. Call for synergy in the work plan. TDWG coordinated.

Graham: could broaden the perspective of genetic resources to the world of physiologists, and others.

GCP Data Quality - Improvement and Assurance (Thomas Metz)

- Rapporteur: Richard Bruskiwich

Introduction (Thomas)

Vision

2005 - project 28 - improvement in quality of existing db's.

- Start community of practice in LIMS
- Implementation of agreed data models in local databases
- Baseline survey on data quality
- Quality assurance strategy

2006

- Community of practice, many small outputs (unpublished experiences; "solutions that work")
 - digital data capture with e.g. bar code in field, etc.
 - digital data capture from instruments e.g. electronic scales
 - Protocols, focused on drought
- Genotyping data quality workshop => best practices for genotyping, data checking and integration

Reality

- 2006 approach lacklustre
- => thomas material from notes...

IRRI system

- shared with CIMMYT

ICRISAT

- to submit software to CropForge

CIP

- barcoding
- statistical controls
- digital (WFI) capture

CIMMYT

- use of handheld PDA's in data collection and archiving

Delivery behind schedule, No cost extension request envisioned

Aim: community of practice, based on remote collaboration Conclusion: difficult to achieve, maybe won't be achieved until end of project

Genotype & Data Quality workshop

- very little initial interest months ago
- changed now that lots of data, DQ issues apparent, integration problematic

Overall conclusion:

- too many, too small deliverables - low funding disincentive to delivery
- face-to-face meetings also needed

Discussion

- Unfinished business
 - Workshop to happen (IRRI)
 - To finish small deliverables

Theo: background of project

- Felt that DQ increasing issue
- No idea how to tackle it
- Brainstorm to identify strategy
- Whole spectrum, from capture to statistical validation

Fred:

- Case study: molecular marker profiles
- Did for number of crops, number of markers => file contents needed to be checked (QC)
- Not sure if this was documented

Thomas:

- workshop of practitioners across crops & disciplines/experience

Fred:

- Data checking of files (e.g. genomics)

Anthony:

- First principles:
- Established common workshop - works - but not heterogeneous
- Documentation alone won't help

Theo:

- automatic procedures imposed on the templates
- could go a step further

Samy:

- GPGs funding data quality
- CSI - GIS coordinates (Isaiah Mukema, IRR)
- Synergy with other proposal

•Theo:

- will investigate opportunities

Thomas:

- target data quality @ data production

Theo:

- Dave:
 - Example audits of the process is recommended.
 - Before molecular sets, DQ analysis not completely possible

- Number of the Institutes have experience globally.
- Need to assess probability of errors occurring

Graham:

- baseline audit in 2005: answer is "everything fine"

fred:

- major risks - end users not auditing files with too much automation.

Dave:

- Documenting what can go wrong in the process.

Fred:

- large genotypes independent of large phenotype

Theo:

- but Thomas is having a difficult time with this.
- surprise at this ARM at number of examples of DQ concerns

Dave:

- don't forget IP
- build up a repertoire of use case of what can go wrong?

Reinhard:

- need to know DQ required for "fitness for purpose"; analyses may be tolerant

Graham:

- how to reorganize task for efficacy?
- how do we document "bitter experience"

Wrap Up: Where do we go from here?

Thomas:

- Workshop?

Theo:

- Do we know who to invite?

Thomas:

- inviting primary users of the data

Theo:

- not funding LIMS, but ICRISAT/BECA experience could be shared

Thomas:

- One workshop could be ICRISAT/BECA => COP
- Possible workshop on WUR/CGN ISO experience

GCP Platform Development (Graham McLaren)

- Rapporteur: Thomas Metz

Introduction

Graham gave a [presentation](#) on progress in the Platform Development Project concluding that the platform foundation was now well established and alpha versions of products from most subtasks are available. However the expected deliverable for this ARM was a platform with working beta versions of applications and datasources together with external critical assessment by users. The opportunity to launch the platform at this ARM has been missed and this will have some severe consequences for future planning. However, the agreed revised target is to have the beta deliverables and critical assessment for all subtasks by the end of 2006 and to present the platform to a GCP side meeting at PAG in January 2007.

Questions/comments during presentation

- GCP Platform Compliance
 - Samy: Perl should be considered in next year's workplan (to be discussed in 2007 workplan section)
- Intellectual property issues with DIVA-GIS and scalability of some applications
 - D. Marshall noted that the use of Google Maps has some licensing

implications - it is free for private use but may not be free or publishing geographic data. This needs to be investigated further

- He also asked whether CMAP is scalable and noted that alternatives are being developed in his group.

Workplan for the rest of 2006

Revised Workplan (Overview)

- Alpha-versions need to be released as beta-user versions (accessible to anybody with beta level documentation)
- Users will be selected for each platform application and asked to provide feedback and critical assessment
- The beta platform will be presented to the GCP community at a side meeting of PAG in January 2007 (Guy Davenport agreed to help arrange this meeting)

Discussion

- The platform needs to accommodate standards for resources outside the GCP (BioCase). However delivering a working platform must be the priority (Response by Graham).
- Speed of data access may be a problem with Web Services. Caching solutions for storing data and intermediate analysis results in XML could be investigated. Applications requiring rapid data access will probably use direct JDBC connections rather than Web Services, and caching data subsets extracted via Web Services as Platform Compliant local data sources might be one strategy.
- Next year we must implement the strategy of short cycles of user feedback and use case presentation
- We must publicize tools as soon as they become available, reference to use cases, GCP newsletter.
- We should investigate opportunities to out-source development tasks. Graham remarked that this was not usually possible because of the rapid evolution in design of most applications which makes firm specification for outsourcing difficult.
- Richard suggested that more collaborative coding (pairwise coding) might

be a way to speed up development and there was some discussion on tools available to help achieve this.

- Genotyping + passport data should be delivered with adequate tools: Genomedia task, Koios are high priority and very important (D. Marshall)

Detailed Workplan by Sub-task for completion in December 2006

- GCP platform middleware: framework and components – M.Senger and all partners.
 - Martin agreed to compile a list of all Platform Compliant Datasources and to work towards an application to check compliance of such datasources.
 - Samy agreed to extend that list to GCP Data compliant sources.
- Web Service/Internet Data Source Integration - Martin Senger.
 - Martin to complete the integration of GDPC Browser as a GCP Platform compliant application
 - Graham to test and provide user feedback
- Generalized query engine and result integrator - Richard Bruskiwich.
 - Richard agreed to make Koios available as a beta version and get a critical review from Hei Leung
- Germplasm Genotype visualization tools- Akinnola Akintunde.
 - Graham to check status of pedigree viewer and agree revised targets with Akin.
- Hibernate adaptors to the ICRIS database - Jayashree Balaji
 - Martin to assist Jayshree in verifying GCP Platform compliance of ICRIS for genotyping and passport data.
 - Jayashree to deploy some applications (Genomediun, DIVA-GIS and/or Koios) from other sub-tasks and obtain some user feedback.
- GenoMedium interface for genotype, QTL and map data analysis and visualisation - Guy Davenport.
 - Guy to verify that GenoMedium queries can be passed to Structure, Tassel, CMTV, GDPC, etc.
 - Guy to get Marilyn to review Genomedium by end-year

- Genomic Sequence analysis and visualization - Manuel Ruiz.
 - Manuel to document user feedback GeneMapper Web interface
 - Richard to organize user assessment of this application at IRRI.
- Geographical data analysis and visualization - Reinhard Simon.
 - DIVA-GIS + Structure, genotyping and passport data available as beta applications.
 - Test and review by users: Mark Ghislain, Brigitte + Marilyn
- Gene expression data analysis and visualization - Masaru Takeya.
 - Masaru to document user assessment of the Cis-Element data mining tool.
 - Masaru to obtain input from users on which data mining tools should be added into MaxD.

Ideas for 2007 Project Proposal

How do we involve users more in 2007

- Given that we have missed the ARM opportunity, how do we get user feedback and input into our 2007 workplan?
 - monitor user satisfaction (Samy)
 - use the helpdesk (2007)
 - PAG used for GCP closed meeting for platform demonstration/feedback (Guy + Graham to organize)
- These are useful tools for the future, however the timeline of october 20th 2006 for 2007 proposals means that they cannot have much impact for this year. Hence the following actions will be taken:
 - Developers proposing sub-tasks for 2007 will be required to have biologist team members and document one or more biological use-cases for which the platform component will be used. This will include an output to demonstrate/document the solution of the specific use-case by ARM 2007.
 - Graham will canvas GCP biologists, particularly some power users asking them to describe bioinformatic workflows for which they require platform solutions (Marilyn, John Bennet, etc..).

Discussion

- Raj noted that the Central Repository should be wrapped as a GCP platform compliant datasource in 2007.
- Samy suggested that the development of a Perl data source interface should be considered in the 2007 workplan.
 - Richard: GCP funded scholar at IRRI is working with Perl and could work on this.
 - Martin: Needs Perl model which requires some resources
 - Graham: Our emphasis for 2007 must be more GCP user visibility and therefore extending the platform infrastructure should take a lower priority than use-case driven application and datasource implementation.

SP4 IP & Legal Issues (Victoria Hensen-Apollonio / Sullivan)

Context

VHA:

- GCP Consortium organisation
 - Contract allows all GCP funded technical property for open usage within consortium
 - Each individual institution has IP assignment in letter of the employment contracts
 - Institution defines what happens when IP exits to the outside (the GCP)
 - Problem is that each institution needs to achieve consensus on how their contributions are to be released outside.
 - Might be good to to put all equivalent policies together side-by-side
 - Recommends GPL-Gnu but... Institutes own TP
 - Default decision implies acceptance of IP
 - Individuals do not have right to give away to open source

Open Source License

Theo:

- GCP contracts could/should provide open source expectations.

Sullivan:

- ...Steering committee has not approved (yet)
- Do have "humanitarian use".
- Introduction of change is problematic.

VHA:

- Could start by an assay of pertinent IP

Theo:

- Open source in workplan proposals
- Workplan into contract
- Sign contract
- ???

Graham:

- Open source written into work plans with details
- Existing code contributions should be contributed?

VHA:

- work plan proposals explicit Open Source
- If including existing institutional open source, need to explicitly bind

Thomas:

- External third party OS?

Sullivan:

- GP License is actually a Copyright

VHA:

- Headers in file, "Copyright <gcp partner>, GPL.. funded by GCP"

Guy:

- mixed software license conflicts

Dave:

- solution is to get permission from software owners

Thomas:

- institutions who own the copyright can release software under more than one license (i.e. dual licensing)

VHA: yes...

Wiki Open Content License

VHA:

- Open versus non-open
 - Non-Open? GCP work is probably fine to share
 - Open? May be need Institutional permission

Thomas:

- Wiki by-passes normal Institutional editorial/publishing prerogatives

Graham:

- Open content wiki ideally needs Institutional approval

Theo:

- Would strongly favor GCP wiki open licensing be written into the GCP contract.

Graham:

- IRRI is moving toward "open content default" review by board

Thomas:

- Restricted access doesn't prevent derivative products

Sullivan:

- Need to worry about keeping trade secrets (commercial purpose) secret...Have to mark confidential.

- Need to develop a professional sense of what to post and not to post - the responsibility lies with you.

VHA:

- Basic decision that individuals need to make is to have to decide what to post or not to post

Theo:

- leave to the scientists???!?

VHA:

- need to carefully consider content.

Theo:

- still hope for a suitable IP statement in the GCP agreement.

VHA:

- that seems to be the way to go.

Theo:

- PSC in November in Washington.

VHA:

- October 21st meeting where I'll get further ideas?

SP4 Communication and Interaction (Theo van Hintum)

- Rapporteur: Raj Sood

Theo

- We did good work last year but we can improve. Some time we lack communication or intense communication.
- It would be important to have one to one communication via Skype , telephone etc. It would be good to have everybody's Skype address. On GCP wiki we can put our Skype address or at least PI's can forward their Skype address to Theo or put it on Wiki.
- Short phone calls would be useful.

- We can use internet2 also to improve communication. Shall we try to have monthly communication via phone? If something urgent going on please do not hesitate to call Theo. Let's try to improve the communication.
- Let's not get bogged down with technical things.

Anthony:

- MSN will go beyond Skype.

Richard:

- How would we do the demo on some one's desk top?

Theo:

- Explore these options and we can look at them later. All PI's can try to get Skype with camera. This is for monthly conversations.

Richard:

- Is there any common agreement to have face to face communication or meeting half through the year?

Graham

- It is necessary for platform task. We talked about having such meeting in Jan in CA next year.

SP4 Project Portfolio (Theo van Hintum)

- Theo, we have to sort out some issues related to Web services task.

Theo went over the DRAFT task list for 2007:

Development of domain models. (Richard)

Implementation of web services in GCP. We will have to buy time for Mathieu and Martin.

Creation and maintenance of templates. (Guy)

Management of the CR. (Tom)

HPC (Antony)

GCP software engineering collaboration platform. (Thomas). There we have received strong questions about how 80K was spent.

Data analysis support for existing projects. There are some short no cost

extensions. And then we may save some money.

Development of integrated GCP platform. Graham. Despite the criticism of not delivering, this task should continue.

GCP data quality improvements. We need to talk about it more with Thomas.

Competitive research of Fred will continue as planned.

Development of plug-ins for GCP platform. Martin's project. It should continue and we need to talk about it. This is a difficult one and is related to platform. But it is important to have PI outside of the GCP consortium.

LD based phenotype analysis. (Fred) At the moment 100K is allocated for this work.

Bioinformatics training is dropped. SP4 help desk is to be implemented. Budget 50K.

Implementation of iMas and its continuation. We received good feedback for this tool.

Training component for iMas.

- It is important to support the use of iMas and promote the software and improve on it. Looking at the possibilities of linking it with the platform.
- It is for low level users. But there are tools that are high end.

Discussion on platform development, documentation and training Graham:

- Query tool could be an option to link it with the platform. Get data from the platform and link it with iMas.

Theo:

- Since it will be used by the NARS and it is important to link it with the platform.

Reinhard:

- What about the training support on the platform?

Graham:

- Support the use of platform by providing helpdesk, manuals etc.

Richard:

- Provide technical documentation related to the platform. Thomas has suggested a wiki for it.

Theo:

- How did you do it with Diva?

Reinhard:

- We developed guidelines. We will have to have some introduction to the new users.

Theo:

- Keep iMas out of this discussion.

Thomas:

- There are visual tools available and by using them we can create some things.

Reinhard:

- It is better to have help desk and create look and feel type of web based help.

Theo:

- We should concentrate this discussion as part of training.

Graham:

- Developing technical documentation as to how to use the platform.

Richard:

- It is a good idea to have annual workshop.

Theo:

- We have to do it per plug-in.

Samy:

- Training is essential part of SP4 activities. Off line and online support. It applies to all software. We could have specific activity with respect to platform activity.

Graham:

- It is premature to launch a big training activity. I would not suggest doing big task on training this year.

Samy:

- Perhaps have some travel grants in your hand is a good idea.

Theo:

- There will be budget for help desk.

Reinhard:

- Involve people to do testing.

Theo:

- Coming year we will keep this as platform task. It is a large task and there is room to move.

Get some power users to play with it to learn and identify points for further improvements. Possible get people to write documents while testing. On top of this we have a help desk which can support the help desk for platform. Samy:

- What about the possibility of collaborating with SP5 to organize training? If we have in mind some training outside?

Theo:

- We could keep this in mind. SP5 aims at outside. We are not concentrating outside for now.

Theo:

- We have to define Milko and Mathieu's work. We have few issues to discuss. Milko's task where it starts and ends and when it becomes Mathieu's task. We have to define Martin's task (plug-in development)

Milko:

- At lunch meeting you were discussing low level and high level MOBY services. I do not understand what it means.
- Where is the boundary of low and high level service?

Richard:

- MOBY is a usable technology and we do not have to do in 2007. Deploy the technology. For Mathieu, to pick few crops and deploy more services and integrating them in the network that is usable in order to services in a sequential manner. May be we can expand the network of ICRISAT.

Theo:

- We needed to get the web services to the partners. Provide training to the partners. For Martin the job was to explore the bioinformatics network and

create the bioinformatics network. For Mathieu to take Musa network and see what we can do with the technology, including workflow and use of TAVARNA. We can use potato as other crop. Would this kind of model work?

Milko:

- We can do some services from one point automatically.

Samy:

- I will leave Mathieu alone to develop services for Musa.

Theo:

- Mathieu what do you think? Does this task give you room to move? Martin will give you back stopping.

Mathieu:

- I am concerned about data.

Richard:

- Explore other crops next door for data.

Theo:

- Develop services for Musa and apply to other crops.

Richard:

- There will be some design issues with respect to TAPIR. Get the issues resolved with TAPIR with respect with MOBY.

Theo:

- Milko you go with your approach. We need to have two project proposals with respect to this proposal. Mathieu need support to write the proposal. With in IPGRI the proposal can be tuned well. Mathieu's proposal should have some academic liberties.

Milko's proposal should indicate deliverables clearly.

Theo:

- Item of Martin's project: We want Martin in GCP as he is a great asset. Template task is important and it should be on the desk of a scientist to get data to the registry. There should not be any problems with the interaction with the scientist.

- We will keep Martin in GCP and explore possibilities with other organizations. We will keep the option that Martin is hired by IRRI.
- How do we define his task? We need to integrate publicly available software in the platform. He should backstop Mathieu.

Richard:

- His position will be spread three ways; Moby, plug-in and his role in the platform. We need his input in the development of the middleware.

Graham:

- This may not be Martin's interest. He may want to see the platform to work with TAVARNA.

Reinhard:

- First we should get the low hanging tools.

Theo:

- We have to open the consortium. I can also write to him to formulate his own work plan. I am putting on the table what he should do.

Graham:

- He needs to work with platform to ensure it works with the tools outside.

Reinhard:

- He can work on helping others and address performance issues.

Theo:

- We can have him as technical person for SP4.

Request to Graham to draft the technical specs, Theo can use in his proposal for Martin.

Theo:

- We had a good meeting. Theo will draft short description of projects and will rely on PI's to work the proposal out.
- Date to finalize proposal is 20th October. Get the proposal out by this date.
- 22 October-17 November: Revision of proposal by RAP and GCP-MT
- 20 November-8 December: Revision of proposal by PI's and SPLs.

- 15 December: Final projects announce.

Closing Remarks:

- It is amazing how this community has grown and I thank you all for your contribution.