WP7 Data Integration & Interoperability

Committee members

Amos Bairoch , chair (SIB)

Michael Ashburner, deputy-chair (University of Cambridge)

Lydie Bougueleret (SIB)

Vincent Breton (CNRS-IN2P3)

Susanna-Assunta Sansone (EMBL-EBI)

http://www.elixir-europe.org/page.php?page=wp7

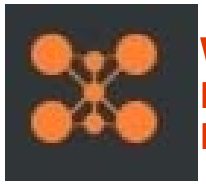
WP 7
Integration &
Interoperability

Interim report - Preliminary work

<<WP7-InterimReport-16Nov2008.doc>>

- Documentation of **existing 'standardization' efforts**
  - of the community databases,
  - of relevant European and international projects

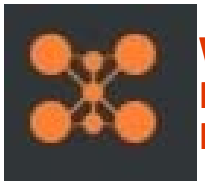
  -> Examples of **databases/tools** implementing these 'standards'
- Identification of **actions** needed
  - to complete, integrate and overcome issues
  - to maximize use of such existing resources
- Development of **strategies** required
  - to overcome the gaps, in line with existing activities
  - to create a consensus set of recommendations and a plan for the adoption of the agreed 'standards'



WP 7
Integration &
Interoperability

Interim report - Four themes

- Programmatic access
  - standardization of the **interoperability technology** to be used to build connections to databases and tools
- Nomenclatures
  - harmonization of **names** and **symbols** of biological objects
- Controlled vocabularies and ontologies
  - harmonization of the **terminologies** used to describe the databases' content
- Reporting requirements
  - standardization of the **minimal information** content to be reported and the **format** used for reporting
    - to guide **deposition** and facilitate **exchange** of the information



WP 7
Integration & Interoperability

Programmatic access - Theme

- Investigate a service-oriented architecture making use of WSs
  - Web Services (WSs) are already widely used both in the bioinformatics and in the grid communities
  - largely promoted by the computing industry
- Leverage on existing projects and recommendations, i.e.:
  - EMBRACE, producing standardized WSs interfaces to molecular databases (EnsEMBL, Hogenom, ProDom, UniProt) and bioinformatics algorithms (BLAST, CLustalW, EMBOSS) to facilitate their integration into biological analysis workflows
    - EMBRACE Service Registry (soon to become: BioCatalogue)
  - BioSapiens, ENFIN, CASIMIR etc.

BioCatalogue

"The Life Science Web Services Registry"

BioCatalogue: providing a curated catalogue of Life Science Web Services.

BioCatalogue will provide a single registration point for Web Service providers and a single search site for scientists and developers.

BioCatalogue will also act as a place where the co
and maintainers of these services.

BioCatalogue will take over the EMBRACE registry

The BioCatalogue team is currently working on its
soon be released for testing to our biocatalogue-fr

"Web Services are hard to find..."

SEARCH

Scientist, tool developers, bioinformaticians will be able to find the right Web Service they were looking for, thanks to an easy and powerful search interface harvesting the information made available by the Web Services providers and the BioCatalogue community.

"My Web Services are not visible..."

REGISTER

Service providers will be able to easily register their Web Services in the BioCatalogue, making them instantly available to the scientific community as well as the tool developers.

"Web Services are poorly described..."

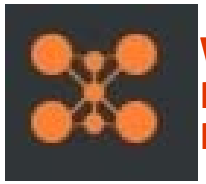
ANNOTATE

Expert curators will provide oversight, monitor the catalogue and provide high quality annotations for services. The wider community will also participate to this effort using social networking for recommending, tagging, commenting and rating the services.

"Web Services are volatile..."

MONITOR

Web Services are volatile. They change their location, capability and interaction or become outdated. BioCatalogue will allow agents to monitor the Web Services and automatically add information to the catalogue.



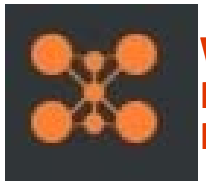
WP 7
Integration & Interoperability

Web services

(preliminary results from the Database Provider Survey)

33. Does your database have Webservices (SOAP, REST, WSDL etc)?			
		Response Percent	Response Count
Yes		30.9%	50
No		**49.4%**	**80**
No; but we plan to introduce it within approximately 12 months		19.8%	32
	answered question		**162**
	skipped question		**2**

Chris Southan, Jan 08



WP 7
Integration & Interoperability

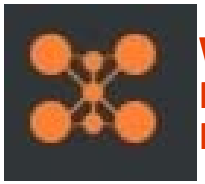
Nomenclatures - Theme

- Encourage pan-organism efforts for gene and protein names
  - Leverage on existing efforts, but promote synergies, i.e.

    - the existing collaboration between the HUGO Gene Nomenclature Committee (HGNC) and the mouse genome informatics database (MGI) to ensure the use of the same symbols in human and mouse in when genes are clearly orthologous

    - the compendium of guidelines nomenclature resource in the framework of the UniProtKB resource

- Enhance taxonomy nomenclature
  - Address species that are not subject to any sequencing effort, therefore not present in NCBI taxonomy database
  - Leverage on global resources, i.e. Encyclopedia of Life
  - Deal with definition of ‘species’ in the light of environmental metagenomics efforts



WP 7
Integration & Interoperability

CVs and ontologies - Theme

- Ensure coordination, leveraging on the existing OBO umbrella
  - 53 are candidate members of the *Foundry*, which ultimately will provide with interoperable, orthogonal, well structured ontologies
  - the *Portal* includes 73 different ontologies (Sep, 2008), of these 33 are the sole or joint products of European groups

_The Open Biomedical Ontologies

Ontologies | Resources | Participate | About

OBO Foundry candidate ontologies

Title	Domain	Prefix	File	Last changed
Amphibian gross anatomy	anatomy	AAO	amphibian_anatomy.obo	2008/06/19
Biological process	biological process	GO	gene_ontology_edit.obo	2008/11/16
C. elegans development	anatomy	WBls	worm_development.obo	2008/01/31
C. elegans gross anatomy	anatomy	WBbt	WBbt.obo	2008/10/29
C. elegans phenotype	phenotype	WBPhenotype	worm_phenotype.obo	2008/11/11
Cell type	anatomy	CL	cell.obo	2008/10/07
Cellular component	anatomy	GO	gene_ontology_edit.obo	2008/11/16
Cereal plant trait	phenotype	TO	plant_trait.obo	2008/04/05
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo	2008/10/29
Common Anatomy Reference Ontology	anatomy	CARO	caro.obo	2007/06/17
Dictyostelium discoideum anatomy	anatomy	DDANAT	dictyostelium_anatomy.obo	2008/05/29
Drosophila development	anatomy	FBdv	fly_development.obo	2007/03/20
Drosophila gross anatomy	anatomy	FBbt	fly_anatomy.obo	2008/06/13
Environment Ontology	environment	ENVO	envo.obo	2008/11/03

WP 7
Integration &
Interoperability

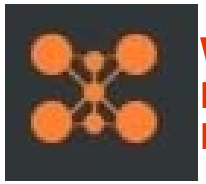
Standards: OBO

(preliminary results from the Database Provider Survey)

35. Does your database conform to specified vocabularies/ontologies ?

	Response Percent	Response Count
No - we are unaware of standards applicable to our data types	**35.6%**	**57**
No - but we plan to within approximately 12 months	11.3%	18
Yes - but not OBO specified	26.9%	43
Yes - we use those specified under the OBO umbrella (www.obofoundry.org/)	26.3%	42
If yes to OBO please list the types view		44
answered question		**160**
skipped question		**4**

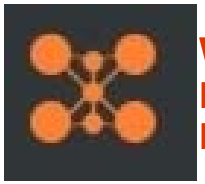
Chris Southan, Jan 08



WP 7
Integration & Interoperability

CVs and ontologies - Theme

- Ensure coordination, leveraging on the existing OBO umbrella
- Address the general funding issue
  - to develop new and maintain existing ontologies
- Focus on domains requiring concerted community efforts
  - disease, anatomy and organismal taxonomies
- Maximize use (of existing) and development of (new) tools
  - to browse, create and edit collaboratively ontologies
- Support new approaches to the problem of annotation
  - wiki-based community annotations efforts (i.e. WikiProtein, WikiGenes)
  - semantic mark-up (i.e. Microsoft Word plugin) and NLP



WP 7
Integration & Interoperability

Reporting requirements - Theme

- Coordinate the development of minimal information requirements
  - leveraging on existing synergistic effort, i.e. MIBBI

MI projects registered with MIBBI

CIMR	Core Information for Metabolomics Reporting
MIABE	Minimal Information About a Bioactive Entity
MIACA	Minimal Information About a Cellular Assay
MIAME	Minimum Information About a Microarray Experiment
MIAME/Env	MIAME / Environmental transcriptomic experiment
MIAME/Nutr	MIAME / Nutrigenomics
MIAME/Plant	MIAME / Plant transcriptomics
MIAME/Tox	MIAME / Toxicogenomics
MIAPA	Minimum Information About a Phylogenetic Analysis
MIAPAR	Minimum Information About a Protein Affinity Reagent
MIAPE	Minimum Information About a Proteomics Experiment
MIARE	Minimum Information About a RNAi Experiment
MIASE	Minimum Information About a Simulation Experiment
MIFlowCyt	Minimum Information for a Flow Cytometry Experiment
MIGen	Minimum Information about a Genotyping Experiment
MIGS	Minimum Information about a Genome Sequence
MIMIx	Minimum Information about a Molecular Interaction Experiment
MIMPP	Minimal Information for Mouse Phenotyping Procedures
MINI	Minimum Information about a Neuroscience Investigation
MIPFE	Minimal Information for Protein Functional Evaluation
MIQAS	Minimal Information for QTLs and Association Studies
MIqPCR	Minimum Information about a quantitative Polymerase Chain Reaction ex
MIRIAM	Minimal Information Required In the Annotation of biochemical Models
MISFISHIE	Minimum Information Specification For In Situ Hybridization and Immunol
STRENDA	Standards for Reporting Enzymology Data

MI projects not registered with MIBBI

MINSEQE	Minimum Information about a high-throughput SeQuencing Experiment
MINIMESS	Minimal Metagenome Sequence Analysis Standard
MIENS	Minimum Information about an ENvironmental Sequence

-> *Portal* includes 28 minimal requirement checklists (Nov, 2008)

consensus view of the essential information on the experimental metadata and associated data that should be reported

-> in the *Foundry* these will be integrated to create interoperable and orthogonal checklists

Projects/MIGS

Minimum Information about a Genome Sequence

1		*General features*
1.1	Domain	Genomics
1.2	Document Type	Primary checklist
1.3	Group	Genomic Standards Consortium
1.4	Main Website	http://gensc.org/
1.5	MI Checklist's Name	Minimum Information about a Genome Sequence
1.6	MI Checklist's Acronym	MIGS
1.7	Current Version Designation	2.0
1.8	Release Date for Current Version	2008-05
1.9	General Comments	Published version available: Stable enough for implementation,

WP 7
Integration & Interoperability

Standards: MIBBI

(preliminary results -160 dbs- from the Database Provider Survey)

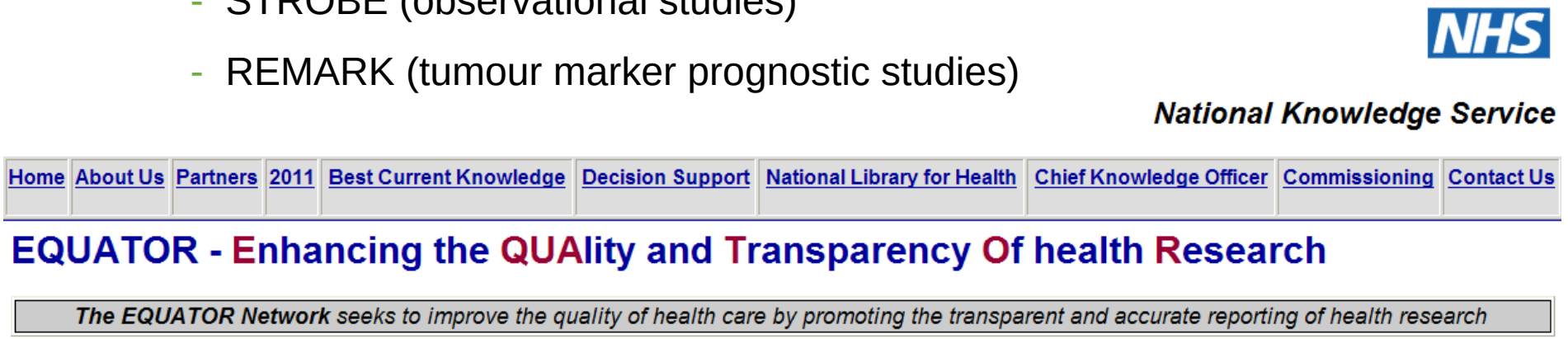
36. Does your database content conform to specified minimum information standards ?

	Response Percent	Response Count
No - we are unaware of any applicable to our data types	**62.3%**	**99**
No - but we plan to within approximately 12 months	7.5%	12
Yes - but not MIBBI specified	21.4%	34
Yes - we use those specified at MIBBI (www.mibbi.org/index.php /MIBBI_portal)	8.8%	14
If yes to MIBBI please list the types view		17
answered question		**159**
skipped question		**5**

Chris Southan, Jan 08

Reporting requirements - Theme

- MIBBI collaboration with EQUATOR network
  - umbrella for minimal information guidelines to report health research, including
    - CONSORT Statement (randomised controlled trials)
    - QUOROM, recently renamed PRISMA (systematic reviews of randomised trials)
    - STARD (diagnostic accuracy studies)
    - STROBE (observational studies)
    - REMARK (tumour marker prognostic studies)

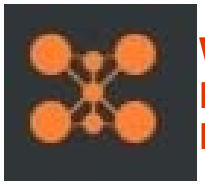


Reporting requirements - Theme

- MIBBI collaboration with EQUATOR network
  - umbrella for minimal information guidelines to report health research, including
    - CONSORT Statement (randomised controlled trials)
    - QUOROM, recently renamed PRISMA (systematic reviews of randomised trials)
    - STARD (diagnostic accuracy studies)
    - STROBE (observational studies)
    - REMARK (tumour marker prognostic studies)

- EQUATOR and MIBBI uptake
  - BioMed Central's journals - with clinical content - now include a link to the EQUATOR and MIBBI in the instructions for authors and peer review guidelines

> For submissions to the journal, all relevant data should be made publicly available either in public repositories or in additional files to be published with the article. Authors should follow reporting and deposition guidelines as summarised by the MIBBI project and the EQUATOR Network, for biological and clinical studies respectively. We are working on guidelines for the formatting and reporting of data.



WP 7
Integration & Interoperability

Reporting requirements - Theme

- Coordinate the development of minimal information requirements
- Encourage pan-domain development of exchange formats
  - variety of file formats, both **tabular** and based on **xml**, focused on particular technologies or particular biologically- or biomedical-delineated community domains
- Synergies to avoid duplication and overcome fragmentation
  - growing number of ‘standards initiatives’:
    - accredited Standards Developing Organizations (SDOs)
    - research community (i.e. GSC, MGED, PSI, MSI) often supported by commercial organizations
  - standards must be interoperable and fit neatly into a jigsaw, with users being able to take the pieces that are relevant to report their study
    - resolve overlaps between domain-specific reporting standards and fill gaps where they exist
    - overcome technical, sociological barriers and funding issue

WP 7
Integration &
Interoperability

Data exchange

(preliminary results -160 dbs- from the Database Provider Survey)

34. Do you exchange data with other databases ? (unidirectional or reciprocal)			
		Response Percent	Response Count
No		31.1%	50
Yes with 1		19.9%	32
Yes with 2-4		30.4%	49
Yes with 5 or more		18.6%	30
	If yes please list the format(s) you use view		99
		answered question	161
		skipped question	3

Chris Southan, Jan 08

Involvement in standards

(preliminary results from the Database Provider Survey)

37. Are you involved in the development of standards ?			
		Response Percent	Response Count
No		57.8%	93
Yes		42.2%	68
	If yes please specify this involvment view		63
	answered question		161
	skipped question		3

Chris Southan, Jan 08

WP 7
Integration &
Interoperability

WP7 next steps

- Continue to engage with the relevant communities
  - A number of WP7 meetings tie in with existing workshops, i.e.:
    - EBI Industry Programme workshop on Disease and Ontologies (org. D Clark) www.ebi.ac.uk/industry/Workshops/workshops.html
    - Set of workshops on synergistic standards and ontologies efforts, including OBO Foundry, MIBBI, co-sponsored by a BBSRC grant (org. S Sansone, P Rocca-Serra) www.ebi.ac.uk/net-project/projects.html#workshop
    - Workshop to advance standards and resources for metabolomics (org. C. Steinbeck, S Sansone) www.elixir-europe.org/page.php?page=metabolomics_workshop
- Report will be extended as the result of closer interaction with
  - other ELIXIR WPs
  - in the light of the results from the ELIXIR surveys
  - several EU and international infrastructure projects
  - related activities in the other ESFRI projects…..
- Final report due in May (last stakeholder meeting in Copenhagen)

EXTRA

Database providers survey

- PubMed: “Database” in title, published in the last 10 years = 5993
  - Mostly clinical dbs (out of scope for ELIXIR)
- As above but top-ten journals with mostly true positives = 1574
  - Nucleic acids research 953, Bioinformatics 246, BMC bioinformatics 114
- As above but filtered by ELIXIR-relevant countries included in affiliation field = 601 (38% of above)
  - Mixed affiliations including outside Europe
  - Includes some advanced publications for 2008 NAR DB issue
- Parsing from the NAR 2008 DB listing gave **410** ELIXIR-relevant (36%) from 1132
  - Journal coverage outside NAR is incomplete
  - Coverage estimate to Oct 2007