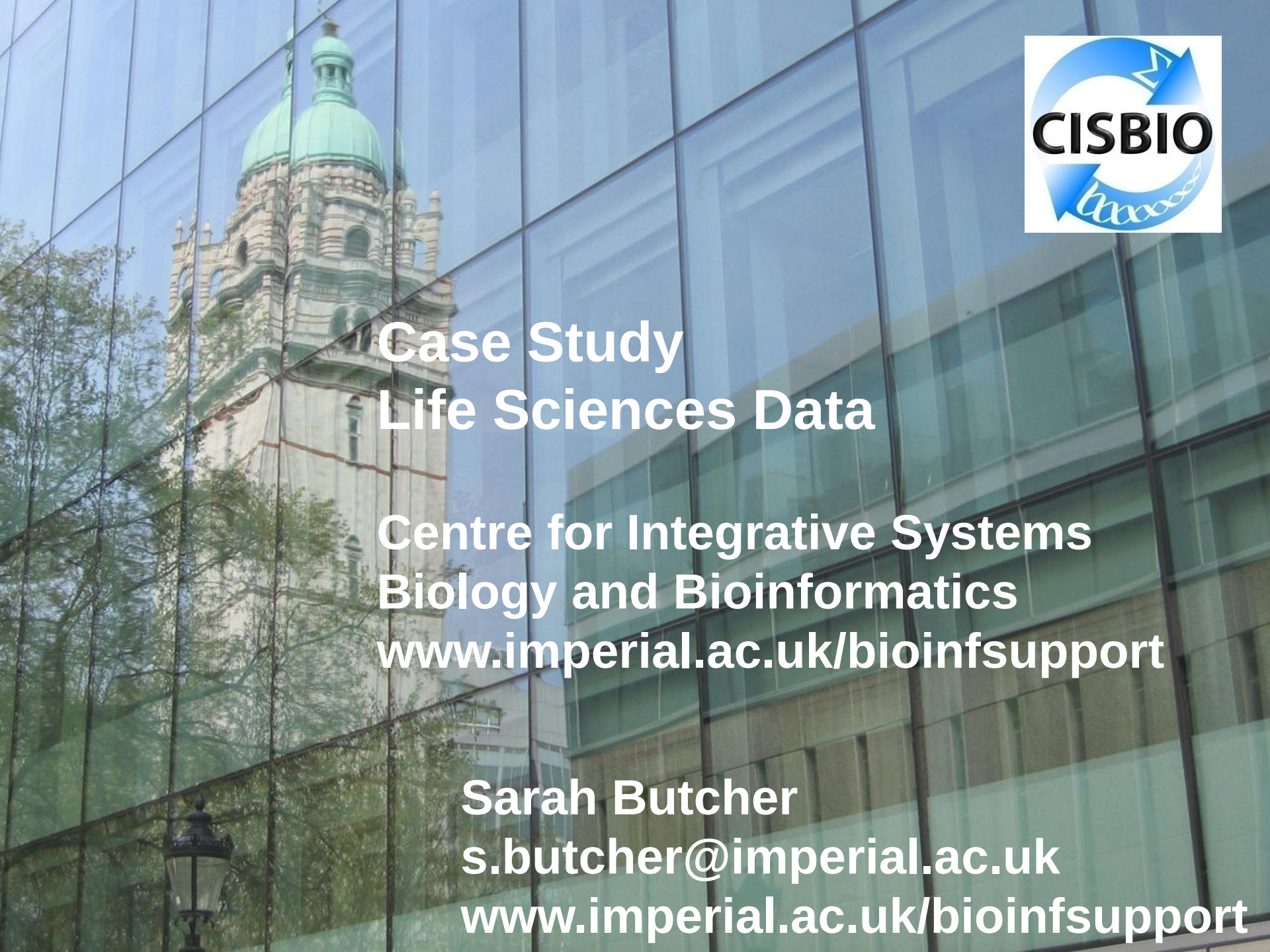


Case Study
Life Sciences Data

Centre for Integrative Systems Biology and Bioinformatics
www.imperial.ac.uk/bioinfsupport

Sarah Butcher
s.butcher@imperial.ac.uk
www.imperial.ac.uk/bioinfsupport

Imperial College London

Bio-data Characteristics

CISBIO

- Lack of structure, rapid growth but not huge volume, **high heterogeneity**
- Multiple file formats, widely differing sizes, acquisition rates
- Considerable manual data collection
- Multiple format changes over data lifetime including production of (evolving) exchange formats
- Huge range of analysis methods, algorithms and software in use with wide ranging computational profiles
- Association with multiple metadata standards and ontologies, some of which are still evolving
- Increasing reference or link to patient data with associated security requirements

Imperial College London

So What Are These Data Anyway?

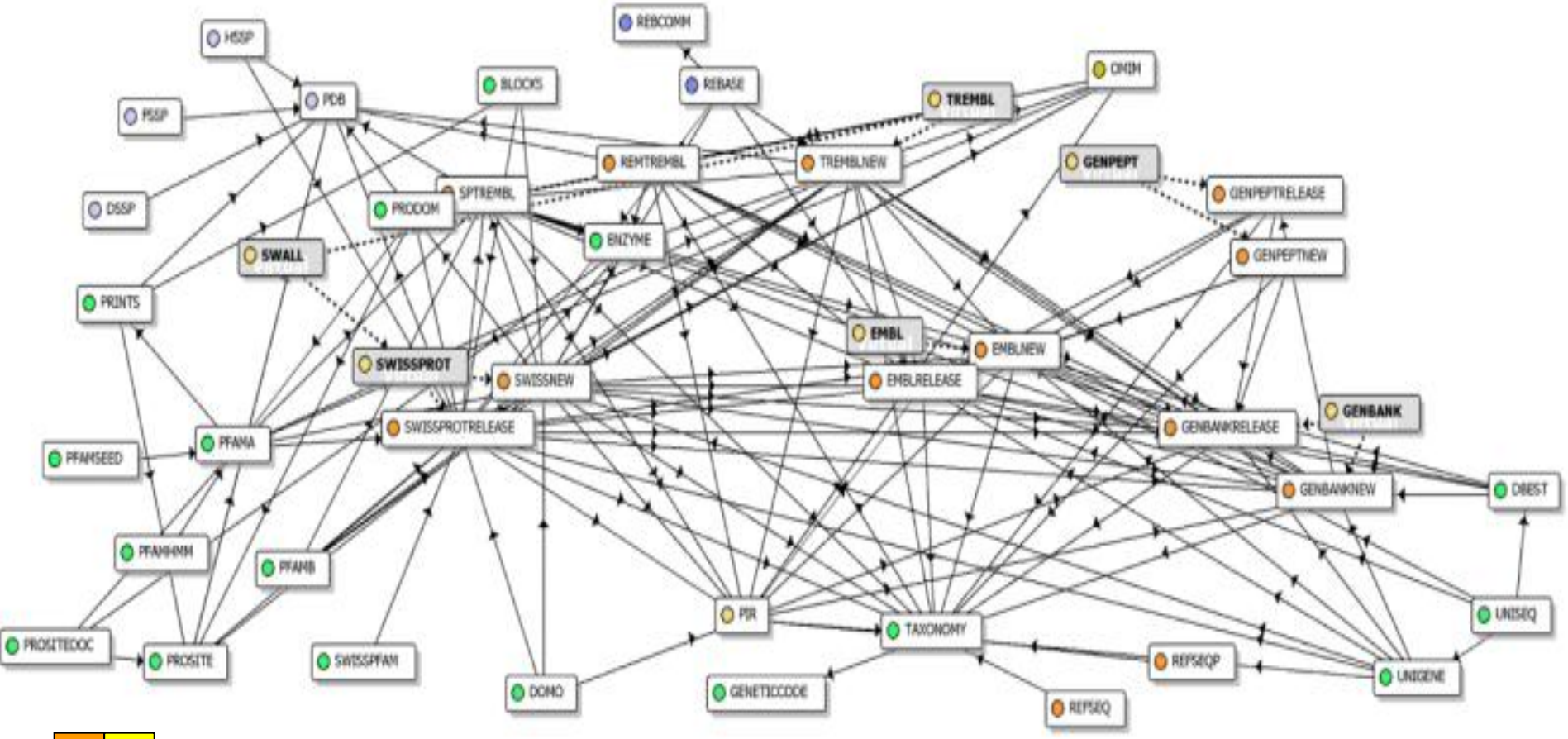
CISBIO

- Raw data files (sometimes)
- Analysed data files (generally)
- Results (multiple formats, often quantitative)
- Mathematics Models (sometimes)
- New hypotheses (hard to encapsulate without context)
- Standard operating procedures (occasionally)
- Software, tools and interfaces (sometimes)
- ‘dark data’ – miscellaneous additional or interim datasets that aren’t directly tied to a publication – might be re-useable if shared and suitable quality

Imperial College London

It's Complicated ….

CISBIO



Primary database – DNA or protein sequence

Secondary - (derived information e.g. protein domains)

Protein structure or other (e.g. crystal coordinates)

Imperial College London

Funders

CISBIO

bbsrc biotechnology and biological sciences research council

EPSRC Pioneering research and skills

MRC Medical Research Council

NERC SCIENCE OF THE ENVIRONMENT

COOPERATION

wellcome trust

THE ROYAL SOCIETY

The Leverhulme Trust

JISC

BILL & MELINDA GATES foundation

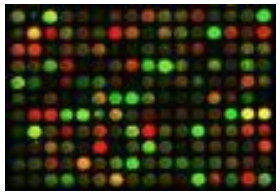
British Heart Foundation

and…. and….

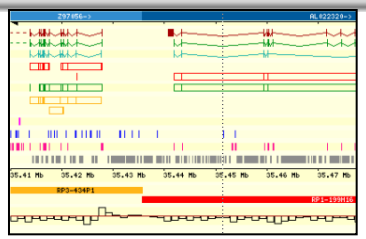
Imperial College London

Many Data Areas

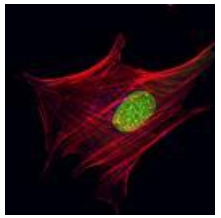
CISBIO



RNA profiles



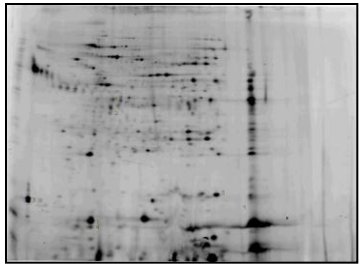
Genome sequencing



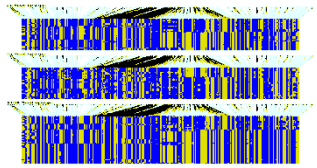
imaging

Improved understanding of complex biological system

Challenges in primary analyses (smaller) AND in meaningful integration (huge)

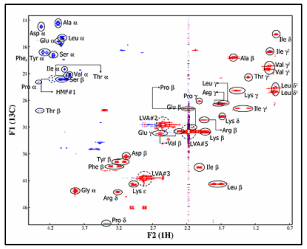


Protein profiles

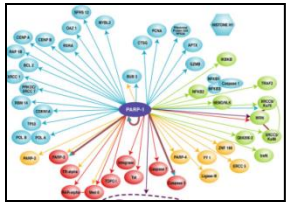


GWAS

Large-scale field studies



Metabolic profiles



Protein interaction studies

Imperial College London

Data Formats

CISBIO

Next generation sequencing - including genome sequencing, re-sequencing and variant detection, RNA-Seq, ChIP-Seq	Binary alignment	BAM	Compressed (binary) version of SAM
	Sequence alignment/map	SAM	Created by alignment programs
	Defining annotation lines on a reference sequence	BED	For visualising annotations in genome browser
	'wiggle' format for continuous-valued data in a track format, also binary compressed version (BigWIG)	WIG BigWIG	e.g. visualisation of GC percent, probability scores, and transcriptome data on genome sequence
	Contains sequence and quality scores	FASTQ	Fasta format sequence and quality data
	Variant calling format (variant positions in genome)	VCF	Text - Often binary format
	Reference-based compression	CRAM	Tuneable binary format for multiple sequences
	General feature format	GFF	Placing features on a genome (reference) sequence

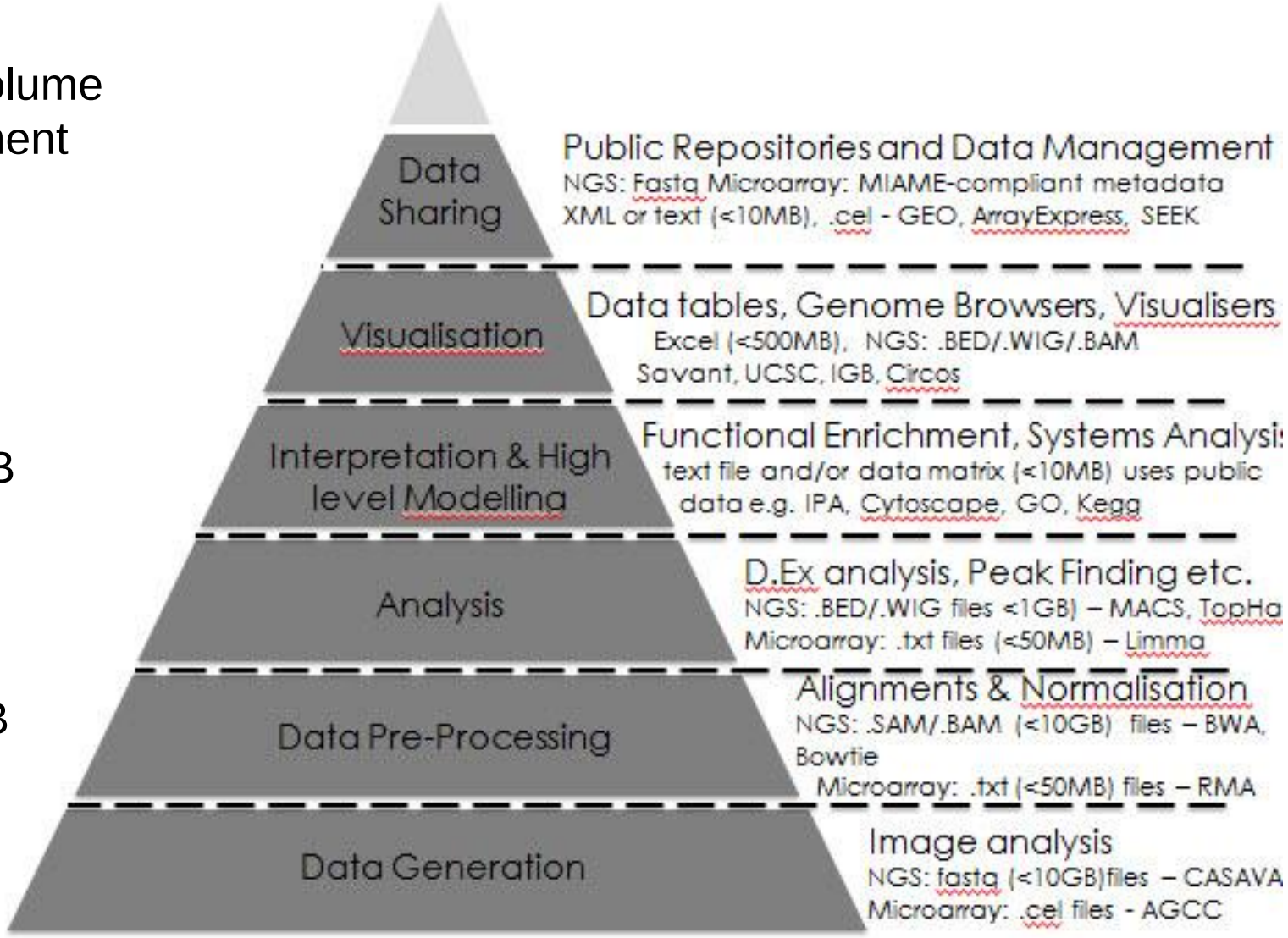
Even for one experimental type, many file formats may be human readable, require require specific software, proprietary or open source….. and Excel spreadsheets

Imperial College London

Data Stages - RNA-Seq Experiment

CISBIO

Total Volume experiment



Imperial College London

Data Lifecycle Management

- **Grant-writing stage** - Data sharing/management Plan
  DMPOnline *https://**dmponline**.dcc.ac.uk/*
- **Pre-publication** - data collection, analysis, storage, some data sharing – *data type-specific repositories, project-based web-sites, metadata annotation, SOP generation*
- **Publication** - publications, submission of data to public repositories, project-based sharing – *journal supplementary materials, public data-repositories, project –specific web-sites, institutional research data catalogs*
- **Post-publication** – updates to active data, data cleansing, archiving of significant, non-changing data

Imperial College London

Data Management/Sharing Plan

CISBIO

- Often the last thing to be written during grant application
- Funding to support plan not always remembered
- Welcome part of grant-planning as encourages focus
- But one size does not fit all
- May not be looked at again until final project phase…..
- Non-generic type-specific EXAMPLES help
- Place for non-traditional ‘data’ often not clear - software, SOPs, models
- Researchers struggle to provide meaningful detail upfront – data volumes, formats, which standards, which repositories – specialist knowledge required

Imperial College London

Public Repositories

CISBIO

- NAR online Molecular Biology Database Collection http://www.oxfordjournals.org/nar/database/c/ currently 1552 databases
- Limited by data domain or origin or both
- One project may require data submission to >1
- May cross-reference data-sets across databases
- Each has its own format and metadata requirements
- Some are manually curated, many are not
- Data submission may be a requirement for journal publication

Imperial College London

(Example)

- http://www.ebi.ac.uk/ena/home since 1980
- Genes, genomes (assembled sequences), raw DNA sequence, annotations
- 3 reporting standards of its own, 5 community-based minimum reporting standards
- Has own XML-based submission system, regularly extended
- Large datasets can take weeks to prepare/validate and generate 100's of thousands of lines of XML, TB of data
- Stable accession numbers

Imperial College London

Bio-Data Standards

CISBIO

- 30+ minimum reporting guidelines for diverse areas of biological and biomedical data
- Few cross experimental types – *confusion, fragmentation*
- Differing levels of use and maturity
- ‘Minimum’ can still be huge – ‘just enough’ movement
- Multiple standard formats for reporting e.g. MAGE-ML
- Not always easy to find associated tools to help use

mibbi

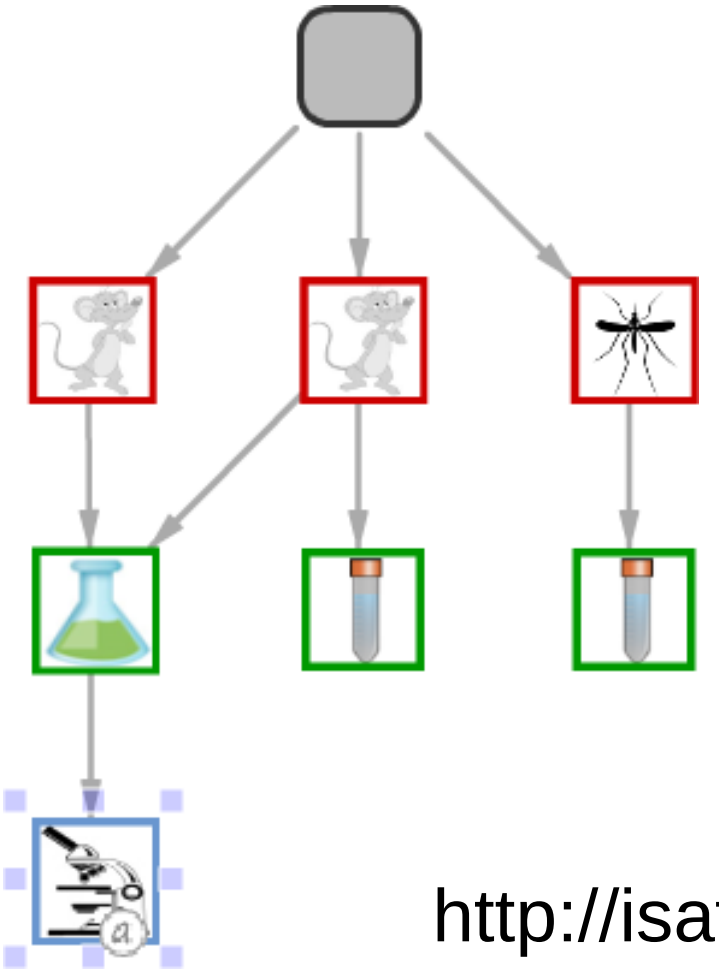
http://mibbi.sourceforge.net/portal.shtml

Imperial College London

isatab

ISA-TAB

CISBIO



- Some do cross boundaries
- ISA-TAB framework – investigation, study, assay hierarchy
- Acts as a framework for associating complex data from a large investigation
- Uptake increasing but still not very widely used

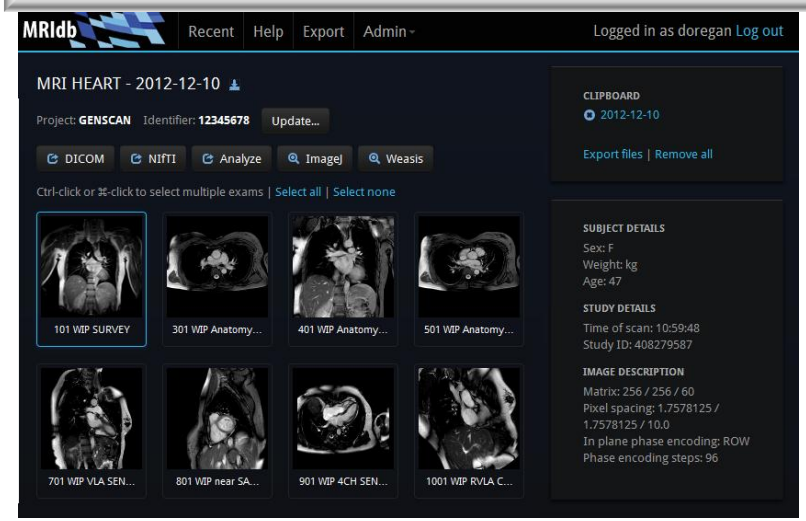
http://isatab.sourceforge.net/format.html

Xperimentr – Tomlinson *et al* (2013) BMC Bioinformatics 10.1186/1471-2105-14-8

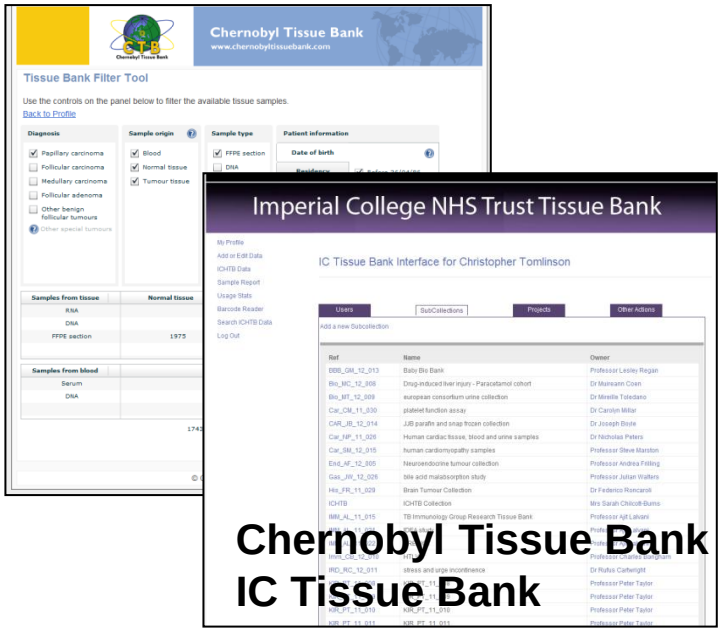
Imperial College London

Specialised Local Repositories

CISBIO

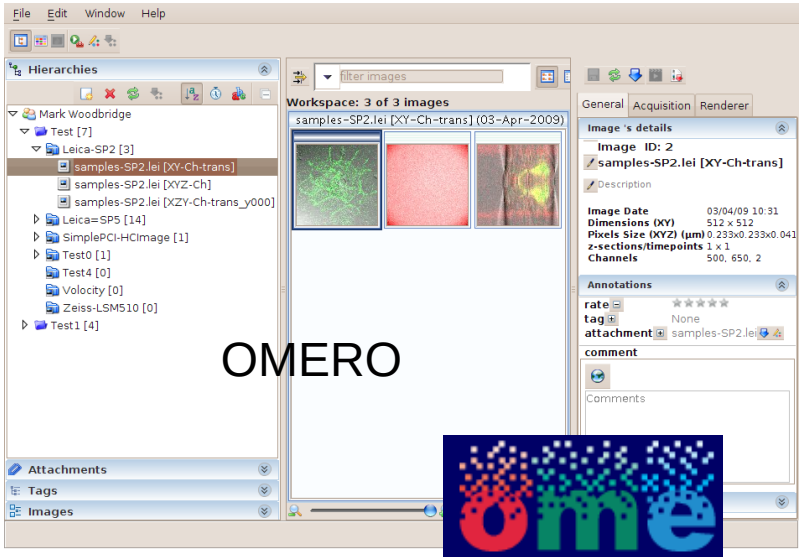


MRIdb

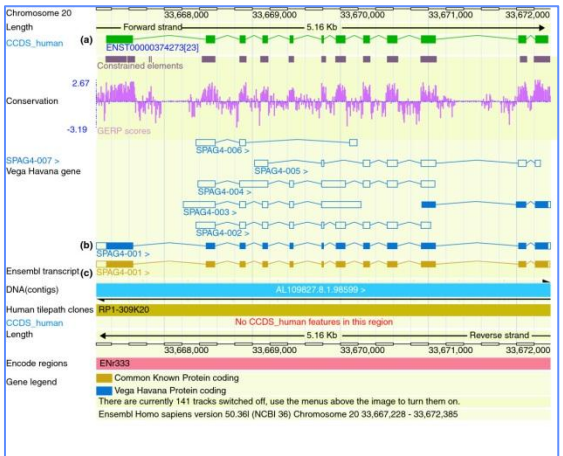


Chernobyl Tissue Bank

IC Tissue Bank



OMERO



e!Ensembl

Imperial College London

Research Outcomes

CISBIO

- Research Councils moving away from grant final report towards reporting research outputs – including publications, datasets, collaborations, impact indications
- Tie to funding information – grant codes, project title
- Ideally, continually updated over project lifetime **and beyond**
- Bring together stable accession numbers for data stored in public repositories, access to 'locally stored dark data', publications, summaries, SOPs, software and tools etc.
- Do need to be findable, stable, searchable, maintained…..

Imperial College London

What to Keep, What to Share ?

CISBIO

- A generic problem area and can be contentious
- Keep all data required to inform a ‘result’
- ‘Worth’ may not be immediately apparent to originator
- Incidental datasets not directly linked to a publication are still getting lost
- ‘Orphan’ datasets with no standard repository - ditto
- Raw data with full metadata are required for re-purposing BUT often huge, requires specialist knowledge, tools to manipulate
- Submission to repositories still requires EFFORT
- Sometimes practical hurdles too great

Imperial College London

Challenges

CISBIO

- Integrative science approaches repeatedly show that complete metadata are vital for optimal data reuse BUT
- Still a complex time-consuming (often resented) task
- Data fragmentation across multiple sites a major barrier to uptake (*can't find it… can't use it…)*
- Practical aspects – research plans change, but data management plans static, cost of storage & curation, difficulty of funding post-project requirements, sheer volume of datasets and complexity of data submission
- Intersection with emergent Institutional general policy

Imperial College London

Resources

CISBIO

- Ten Simple Rules for the Care and Feeding of Scientific Data. Goodman *et al* (2014) PLOS Computational Biology **10** (4) e1003542
- SEEK - http://www.seek4science.org/ (*example open source data repository system)*
- www.blugen.org (*example project-specific web-site)*
- http://www3.imperial.ac.uk/bioinfsupport/projects/systemsbiology/lola *(single project research output example)*
- *OMERO* http://www.openmicroscopy.org/site/products/omero *(example bio-image data management/sharing system)*
- MRIdb: Medical Image management for Biobank Research. Woodbridge *et al* (2013) J Digit. Imaging **26**: 886-890.
- http://www.biomedbridges.eu/news/principles-data-management-and-sharing-european-research-infrastructures

Imperial College London

- Infrastructure Systems Biology Europe http://project.isbe.eu/preparatory-phase/wps/wp2/

- Elixir http://www.elixir-europe.org

- Biomedbridges http://www.biomedbridges.eu/

- Research Data Alliance https://rd-alliance.org/

- Software Sustainability Institute http://www.software.ac.uk/

- Biosharing http://biosharing.org/