

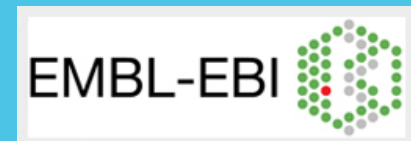


e-infrastructure

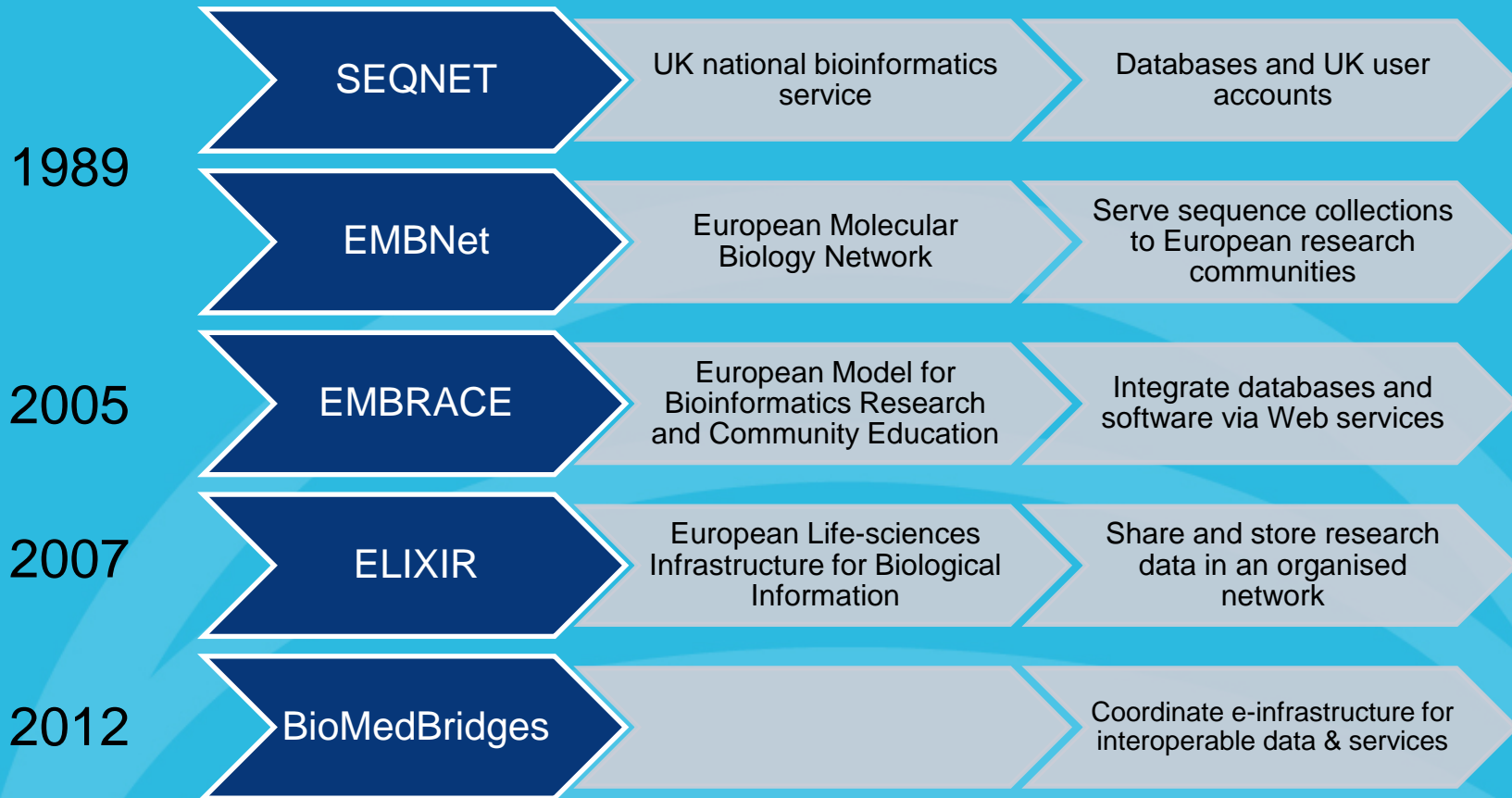
a hackers perspective

Jon Ison, PhD
eIRG Workshop
22-23 May, 2013, Trinity College, Dublin

www.biomedbridges.eu

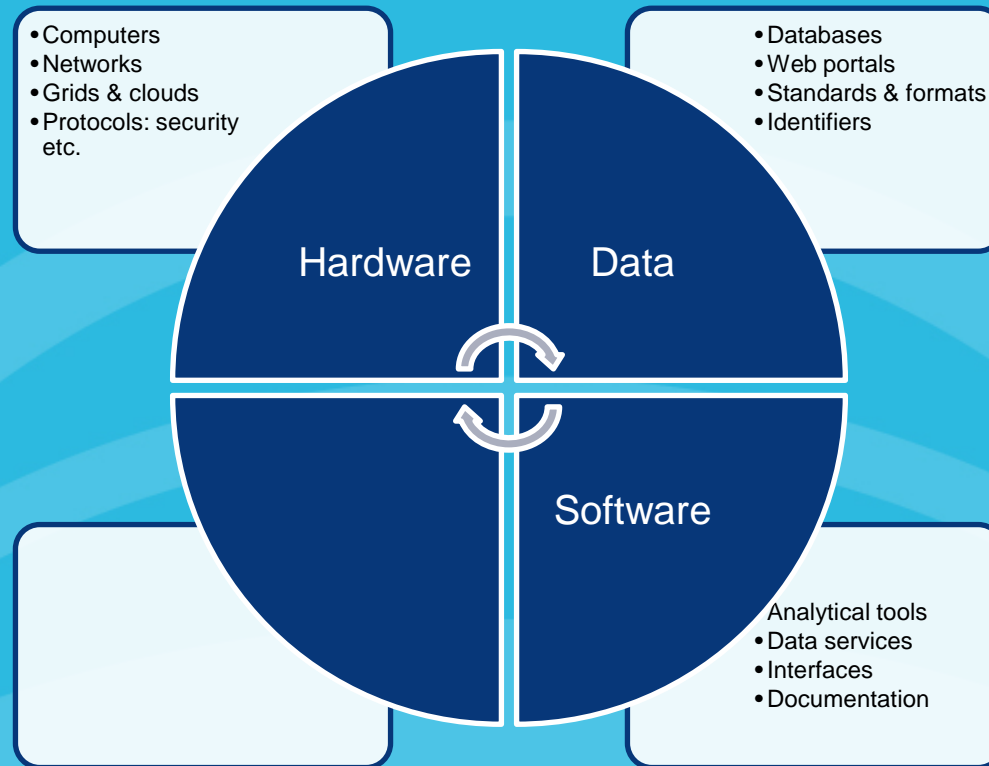


Bioinformatics infrastructures



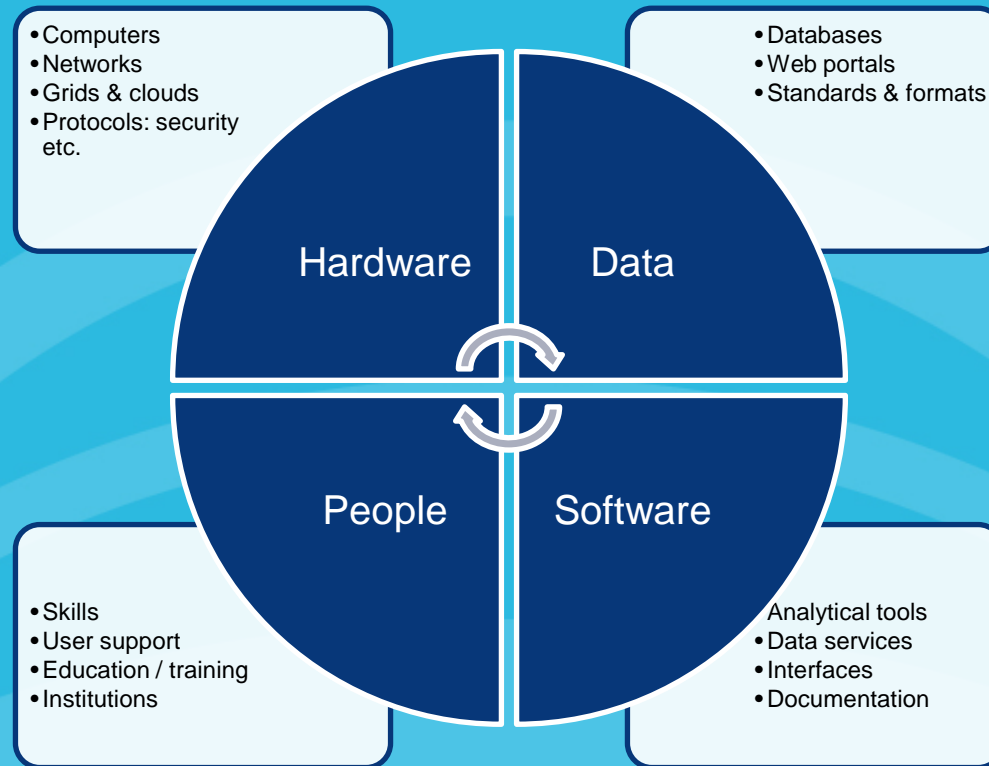
e-infrastructure

1. “The basic facilities, services and installations needed for the functioning of a community”

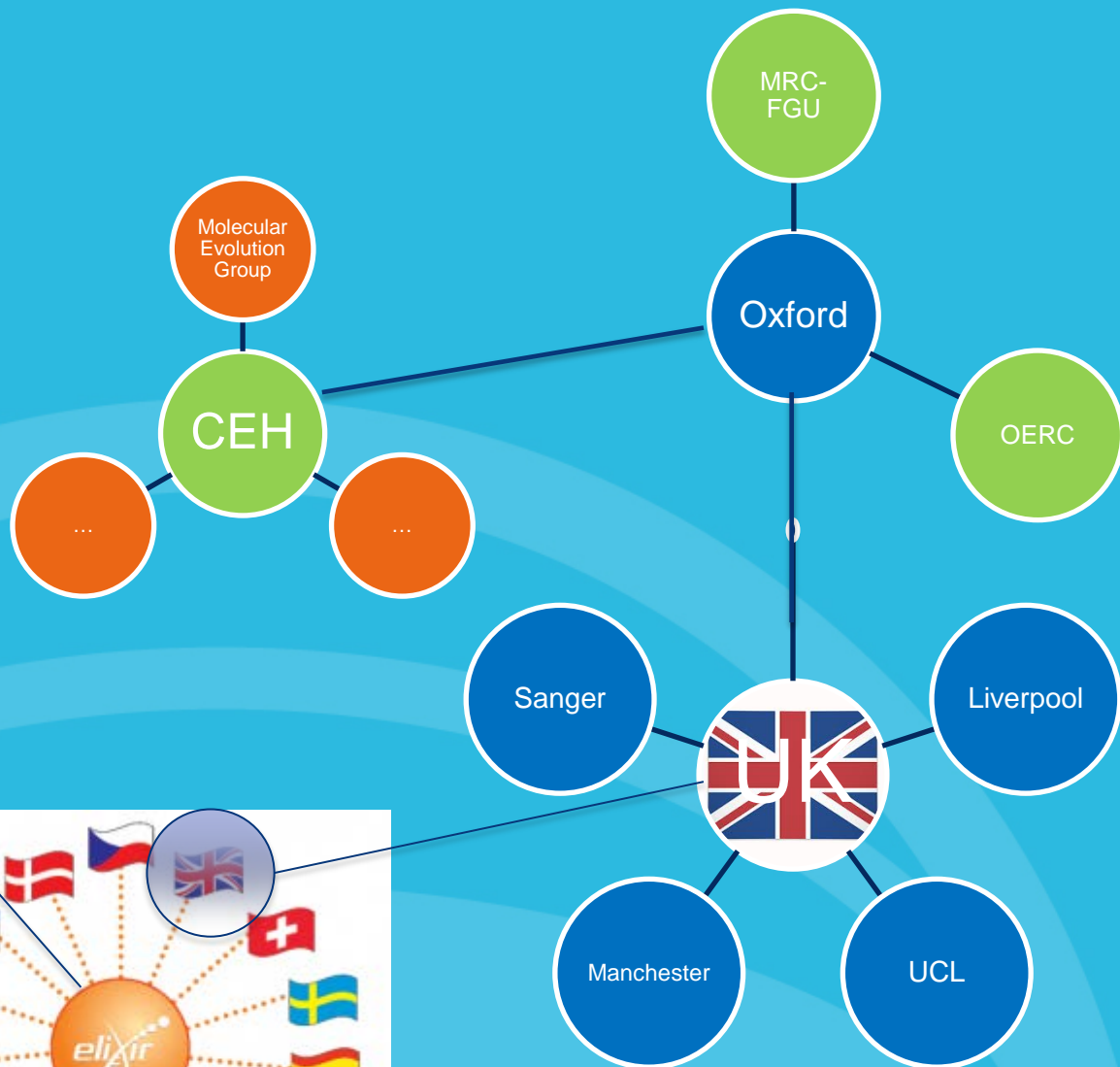


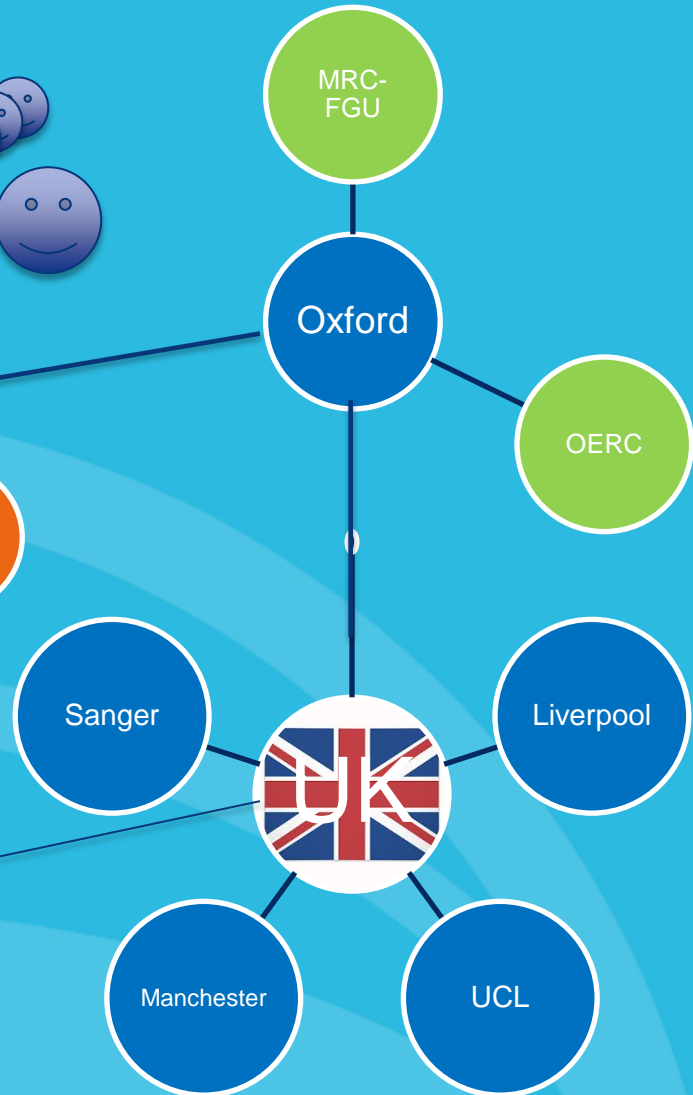
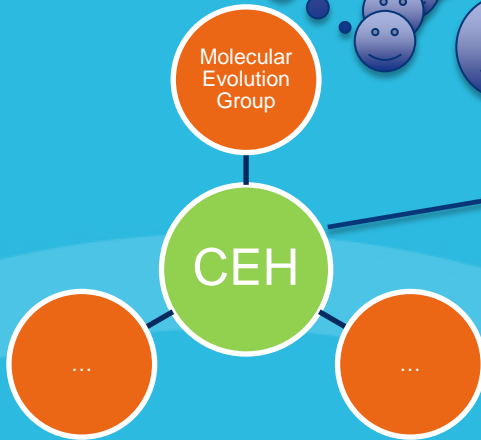
e-infrastructure

1. *“The basic facilities, services and installations needed for the functioning of a community”*



2. *“The underlying base or foundation for an organisation or system”*







Worker bees need ...

Worker bees need ...

£££ to carry on with the R&D they enjoy



Worker bees need ...

£££ to carry on with the R&D they enjoy



£££ to travel and meet like-minded people



Worker bees need ...

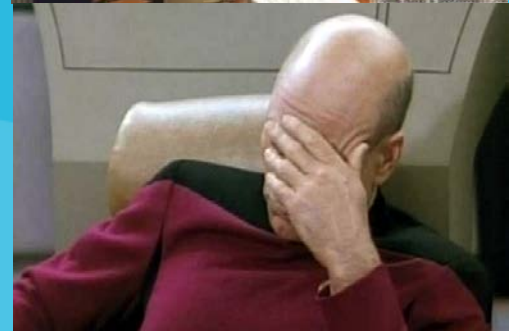
£££ to carry on with the R&D they enjoy



£££ to travel and meet like-minded people



Help with things they'd sooner not be doing



Why bother ?

- Hard to find / understand / compare stuff (tools, data)
- Hard to identify stuff, e.g. tool version, data provenance
- Inconsistent descriptions
- Poor or no documentation
- Lack of examples / sample data
- Too many hacks e.g. file formats and “standards”
- Wasteful reinvention of the wheel

Why bother ?

- Hard to find / understand / compare stuff (tools, data)
- Hard to identify stuff, e.g. tool version, data provenance
- Inconsistent descriptions
- Poor or no documentation
- Lack of examples / sample data
- Too many hacks e.g. file formats and “standards”
- Wasteful reinvention of the wheel

Things *should* be much better!

***Big waste of time
Distraction from science
Can't see the wood for the trees!***



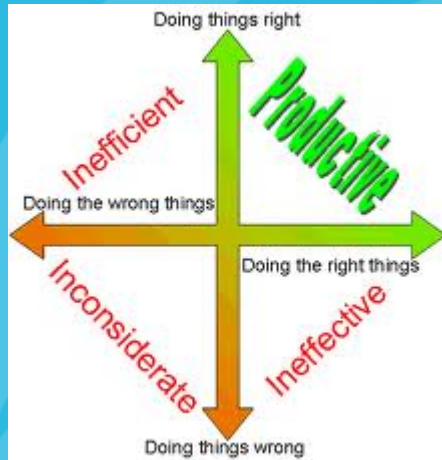
e-infrastructure must deliver ...



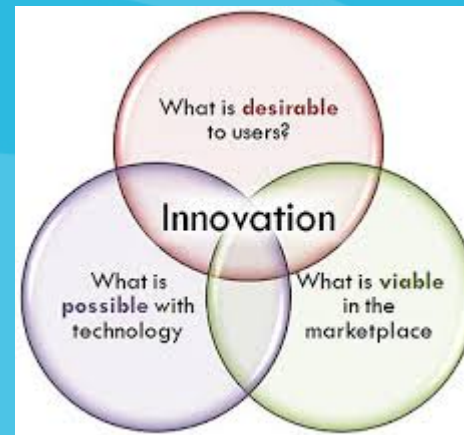
Collaboration – share ideas, resources and work



Efficiency – do more for less



Productivity – do more sooner



Innovation –work smarter, get results!

But get it right ...



Facilitate the worker bees!

- Bottom up: build on good existing efforts
- Small, practical steps

Keep grand ambitions and lofty missions for the funding agencies :)



Data

I want access

The taxpayer paid for it – data must be open

I don't want to log on (life is too short)

Keep my personal data private



Standards & formats

I want data in standard formats

Don't invent new formats

You can't invent a standard (please don't try)

Standards & formats

I want data in standard formats

Don't invent new formats

You can't invent a standard (please don't try)

Tell folk what's out there - encourage them !

Help the worker bees converge on common ground



Software tools

Well documented

Worked examples with sample data

Clean and stable interface

Work as advertised

Supported

Versioned

Open source

wish list



Common identifiers

Resources must be identifiable

Common identifiers allow consistent references to things

Graceful evolution mechanism (to handle change, e.g. software version)

Graceful obsolescence mechanism (so identifiers never vanish)

Plus (just a bit) of provenance



Common vocabularies

Use same terms to describe stuff (data, tools etc.)

- common controlled vocabularies (CVs)

Don't casually invent new CVs – build on existing ones

Beware modelling / ontology construction not an end in itself

Terms with definitions in a simple tree might be enough

Practical steps

Promote best practice

Publish data

Document software for use by others

Use common data standards and formats

Collaborate to build common vocabularies

Collaborate to build common catalogues / registries

- Software tools including data services
- Data resources and datasets
- Common data standards and formats

Talk to others (before starting) to get on the right foot

Don't reinvent the wheel!



Practical steps

Show don't tell - lead by example!

Education and training (workshops, courses etc.)

- users (scientists) and developers
- technical experts

Good documentation

- on all the topics discussed so far
- concise

Be inclusive – engage the community!





What have I done about it?

EDAM – controlled vocabulary for bioinformatics data, formats, identifiers, operations and topics

DRCAT – catalogue of 500+ data resources and services, annotated with EDAM terms

EMBOSS – suite of open-source tools for common bioinformatics tasks, with common interface (EDAM annotated) and standardised documentation

BioMedBridges registry –biomedical software tools

Community efforts – meetings & workshops