e-infrastructure

a hackers perspective

Jon Ison, PhD

eIRG Workshop

22-23 May, 2013, Trinity College, Dublin

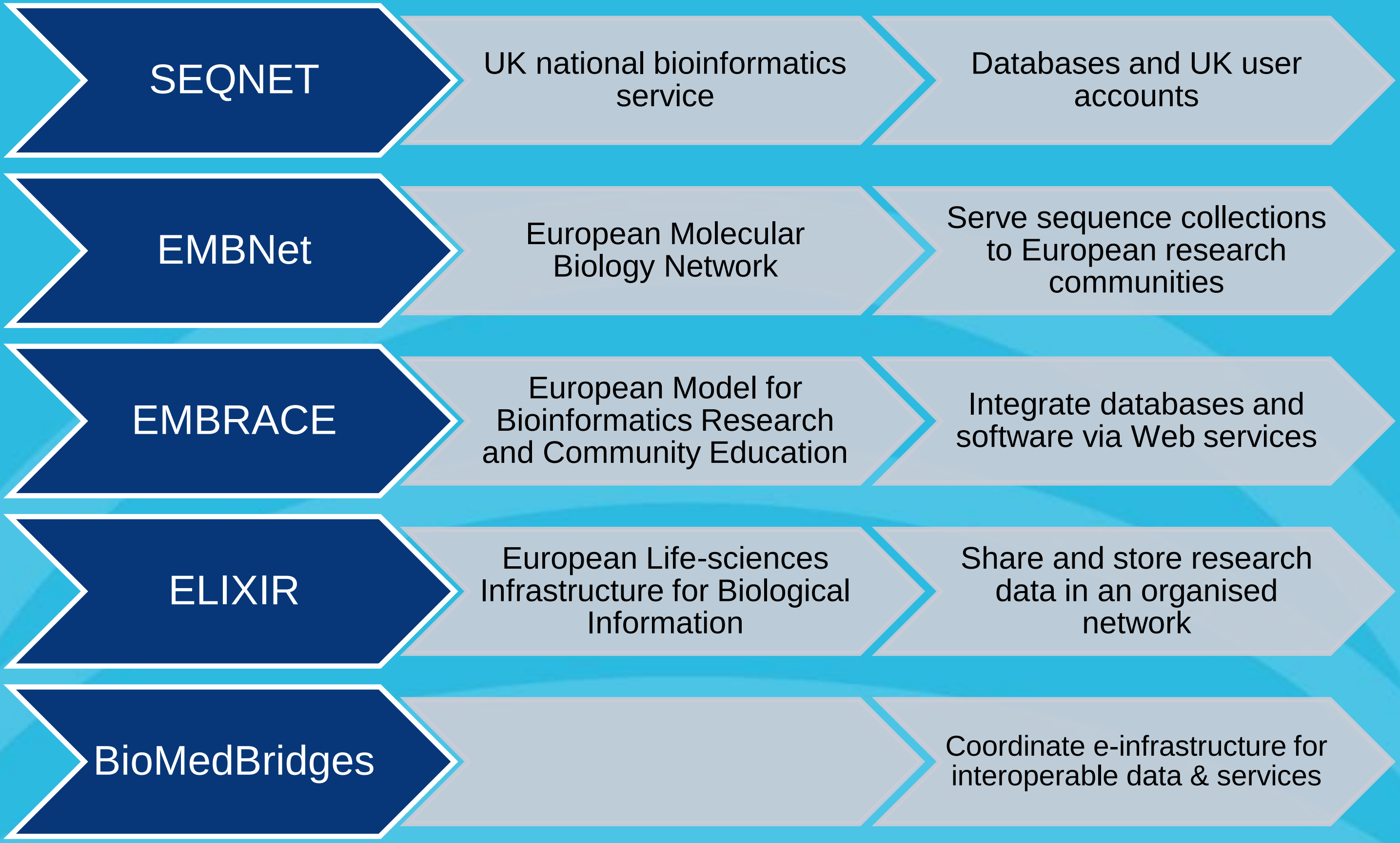
www.biomedbridges.eu

BioMedBridges

EMBL-EBI

SEVENTH FRAMEWORK PROGRAMME

Bioinformatics infrastructures



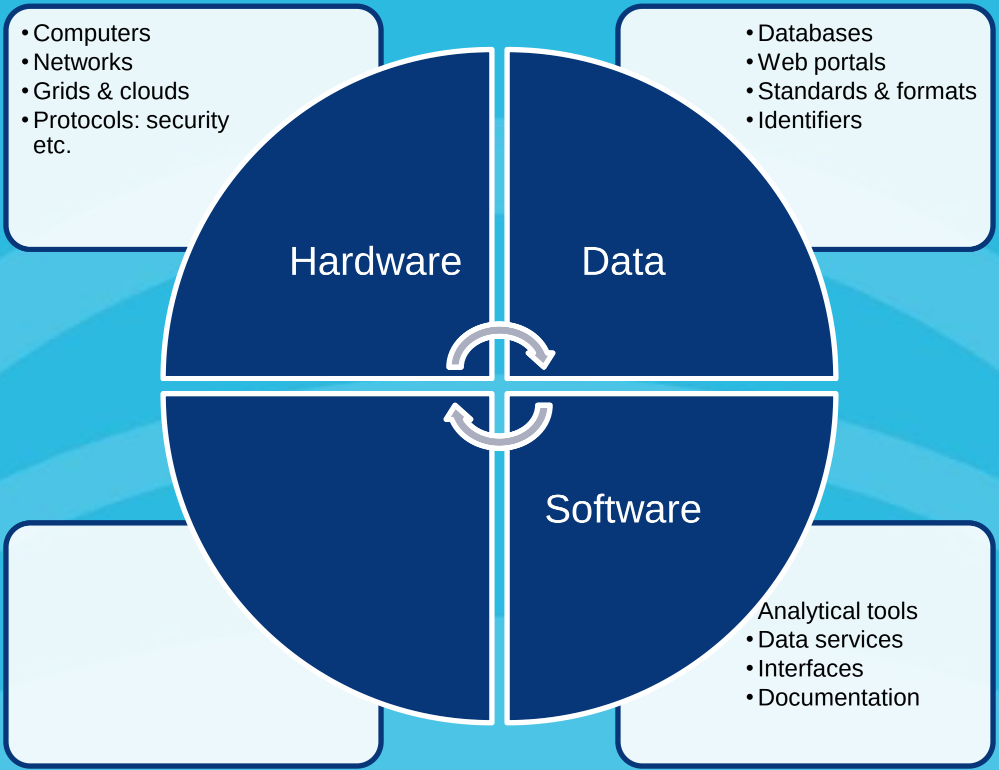
1989	SEQNET	UK national bioinformatics service	Databases and UK user accounts
	EMBNet	European Molecular Biology Network	Serve sequence collections to European research communities
2005	EMBRACE	European Model for Bioinformatics Research and Community Education	Integrate databases and software via Web services
2007	ELIXIR	European Life-sciences Infrastructure for Biological Information	Share and store research data in an organised network
2012	BioMedBridges		Coordinate e-infrastructure for interoperable data & services

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

e-infrastructure

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



- Computers
- Networks
- Grids & clouds
- Protocols: security etc.

- Databases
- Web portals
- Standards & formats
- Identifiers

Hardware

Data

Software

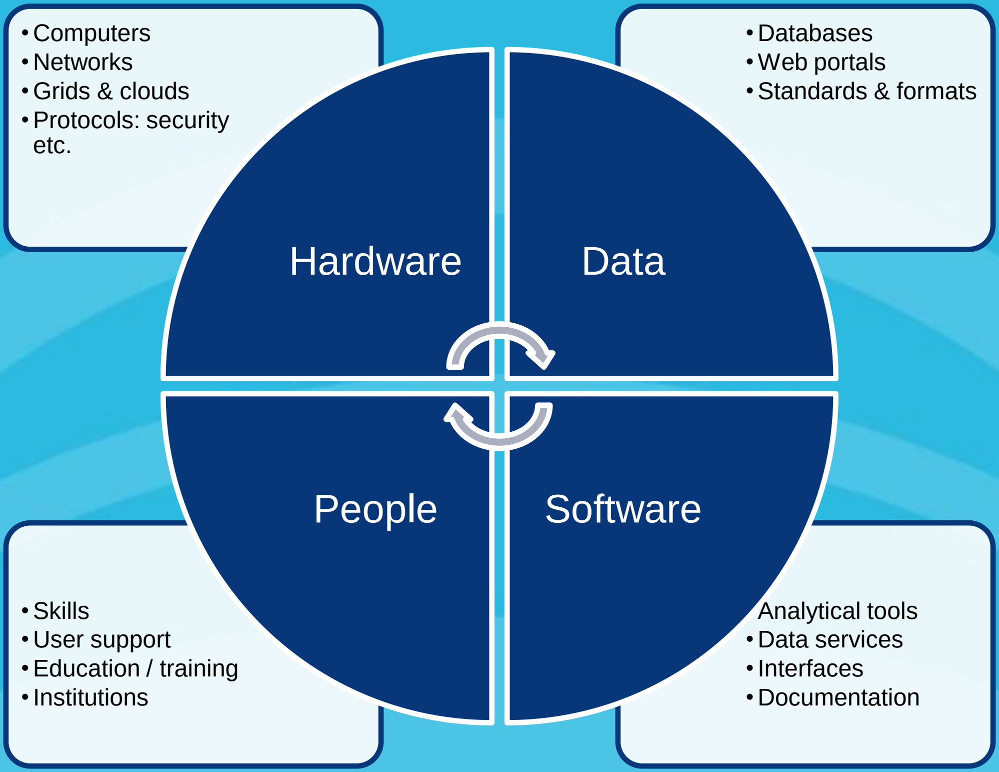
- Analytical tools
- Data services
- Interfaces
- Documentation

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

e-infrastructure

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



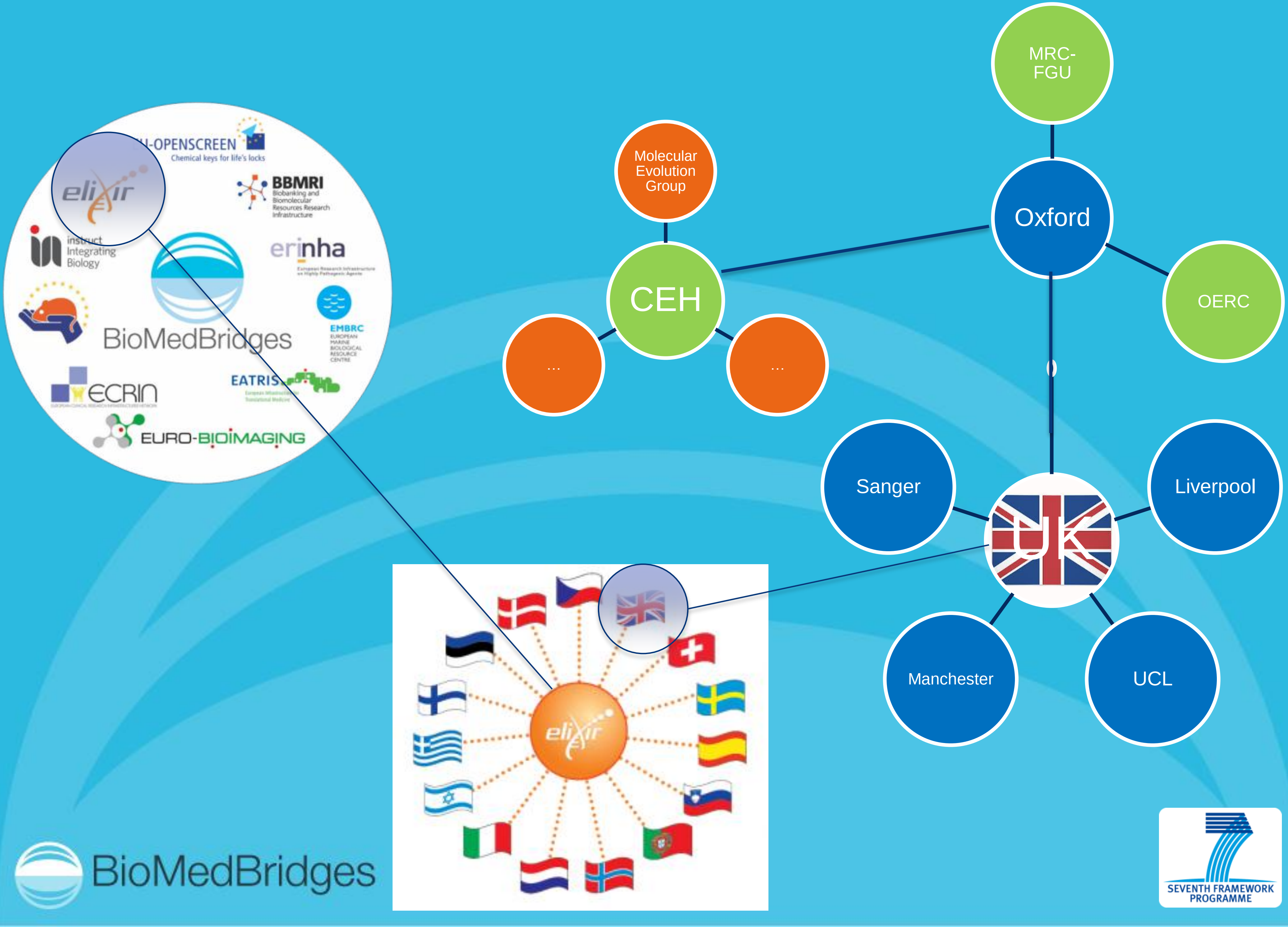
- Computers
- Networks
- Grids & clouds
- Protocols: security etc.

- Databases
- Web portals
- Standards & formats

- Skills
- User support
- Education / training
- Institutions

- Analytical tools
- Data services
- Interfaces
- Documentation

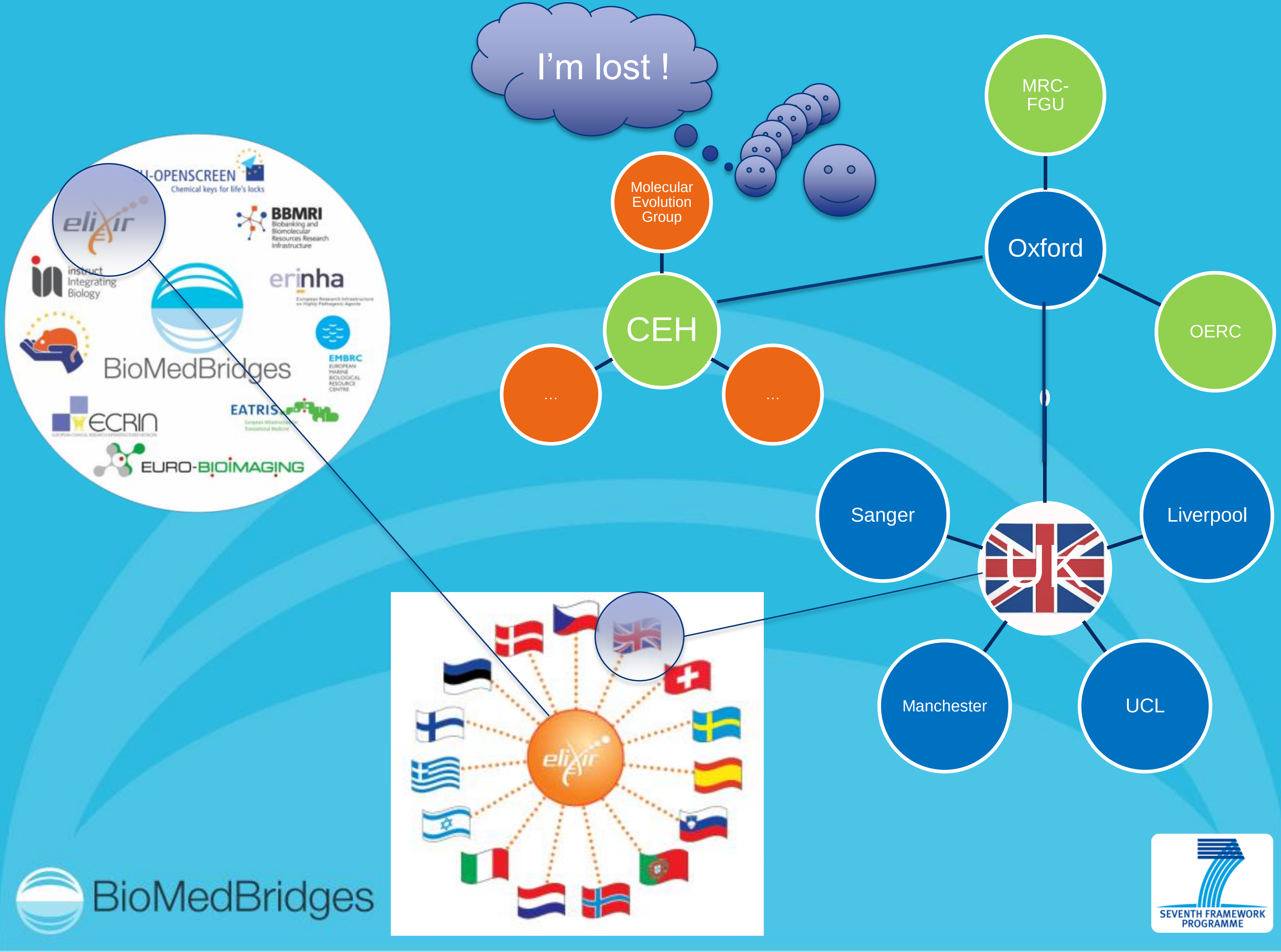
2. "The underlying base or foundation for an organisation or system"



EU-OPENSCREEN
Chemical keys for life's locks
elixir
BBMRI
Biobanking and Biomolecular Resources Research Infrastructure
instruct
Integrating Biology
erinha
European Research Infrastructure on Highly Pathogenic Agents
BioMedBridges
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
ECRIN
EATRIS
European Infrastructure for Translational Medicine
EURO-BIOIMAGING
elixir
MRC-FGU
Molecular Evolution Group
Oxford
CEH
OERC
…
…
Sanger
Liverpool
UK
Manchester
UCL

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME


I'm lost !
MRC-FGU
Molecular Evolution Group
Oxford
CEH
OERC
…
…
Sanger
Liverpool
UK
Manchester
UCL
EU-OPENSCREEN
Chemical keys for life's locks
elixir
BBMRI
Biobanking and Biomolecular Resources Research Infrastructure
instruct
Integrating Biology
erinha
BioMedBridges
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
ECRIN
EATRIS
EURO-BIOIMAGING
BioMedBridges
SEVENTH FRAMEWORK PROGRAMME

Worker bees need …

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Worker bees need …

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

Worker bees need …

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

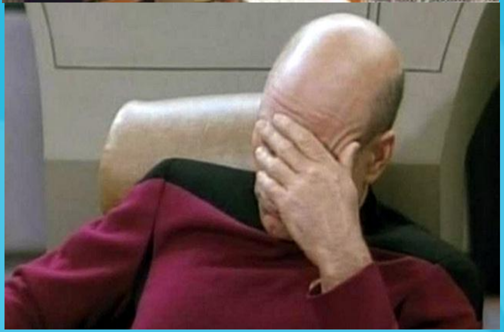
£££ to travel and meet like-minded people

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



BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

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

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

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!

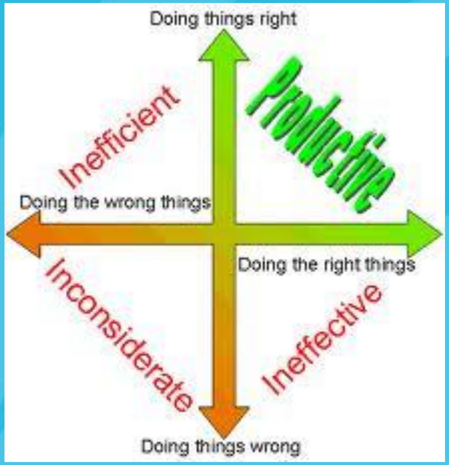
Big waste of time
Distraction from science
Can’t see the wood for the trees!

e-infrastructure must deliver …

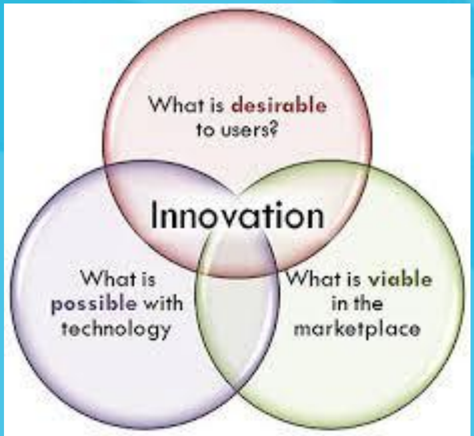
Collaboration – *share ideas, resources and work*

MY TEAM IS SO EFFICIENT, BECAUSE EMPLOYEES' NEEDS ARE MET ON-SITE.
WHEN THE TEAM IS HUNGRY, WE CATER. IF WE NEED TO WORK OUT, WE'VE GOT A GYM. IF WE NEED ENTERTAINMENT, WE HAVE PING PONG.
EFFICIENCY
TED TRIED TO EXPLAIN TO HIS WIFE WHY HE SLEPT WITH HIS SECRETARY.

Efficiency – *do more for less*



Productivity – *do more sooner*


What is desirable to users?
Innovation
What is possible with technology
What is viable in the marketplace

Innovation –*work smarter, get results!*

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

But get it right …

THIS WILL GO A HUNDRED TIMES FASTER!
100
YEAH IN THE WRONG DIRECTION...

© Seppo Leinonen 2006
Pohi Kuokka

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

Common identifiers

Resources must be identifiable

Common identifiers allow consistent references to things

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

Graceful obsoletion 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

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

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