Science Europe:

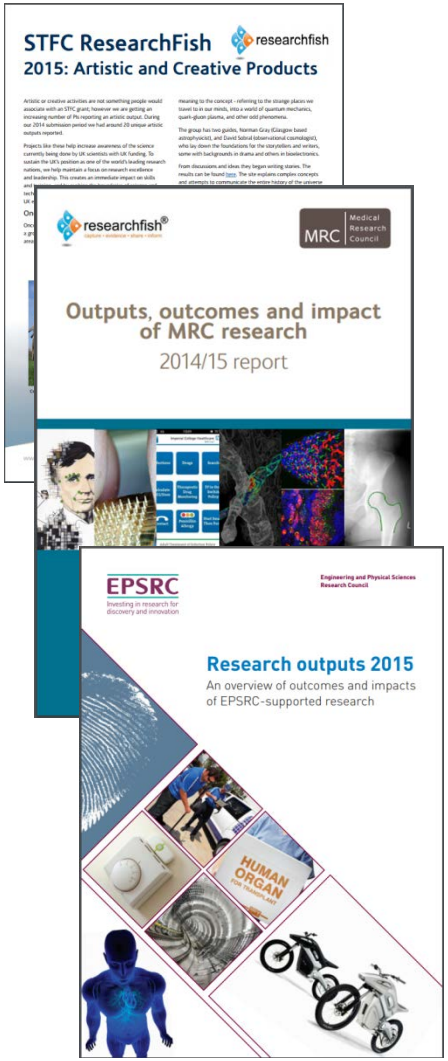
Enhancing interoperability of research activity data

openevaluation
Vienna 2016

Dr Emily Gale
Medical Research Council, UK

emily.gale@headoffice.mrc.ac.uk

Why we collect research information



- **Advocacy**
- Helpful in justifying additional funding for research and explaining research impact

- **Accountability**
- Evidence of delivery to Government
- Answering Parliamentary and other queries

- **Analysis**
- Basis for tracking progress, productivity, quality and impact of research we support, in aggregate, across disciplines
- Data used to better understand impact

- **Efficiency**
- Replacement for final grant reports, updated throughout the lifetime of an award and beyond, understanding time lags etc.
- Common across a large number of funders to build a national dataset on research output
- Providing data to research organisations to re-use

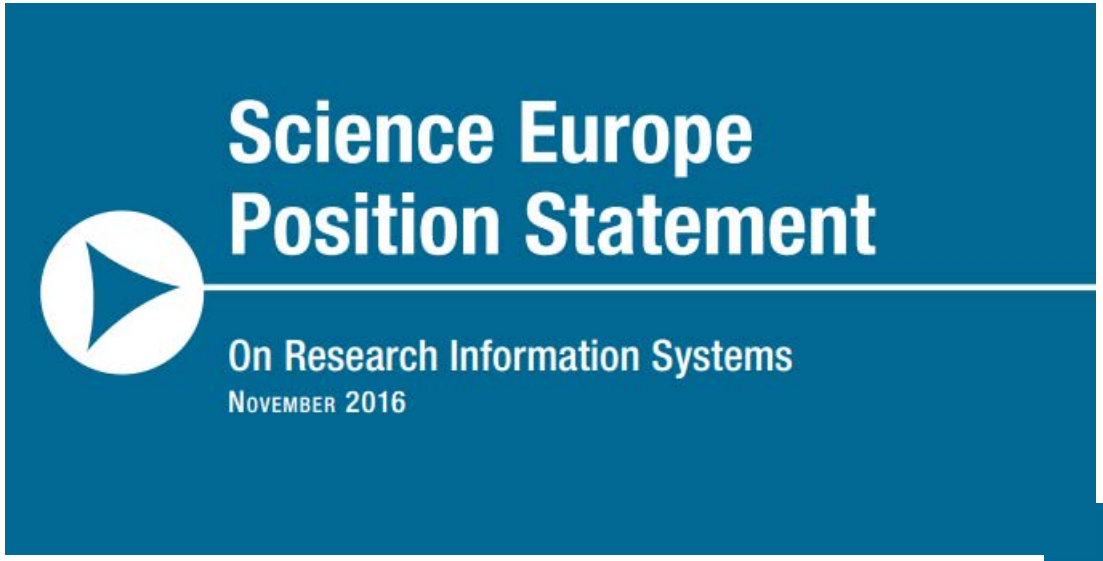
- **Openness**
- Sharing information about research council output via Gateway to Research to business and the public

Research information systems

Research activity data

- **input data:** researchers, research budgets and projects, research institutions, research programmes, etc.

- **outcome data:** scientific publications, collaborations, further funding, IP, products, interventions or other outputs.



Powerful analyses can be derived when different datasets are linked. The reliability, reach and comprehensiveness of these analyses is a function of the quality of the underlying data, their mutual compatibility and their interoperability.

http://www.scienceeurope.org/news/

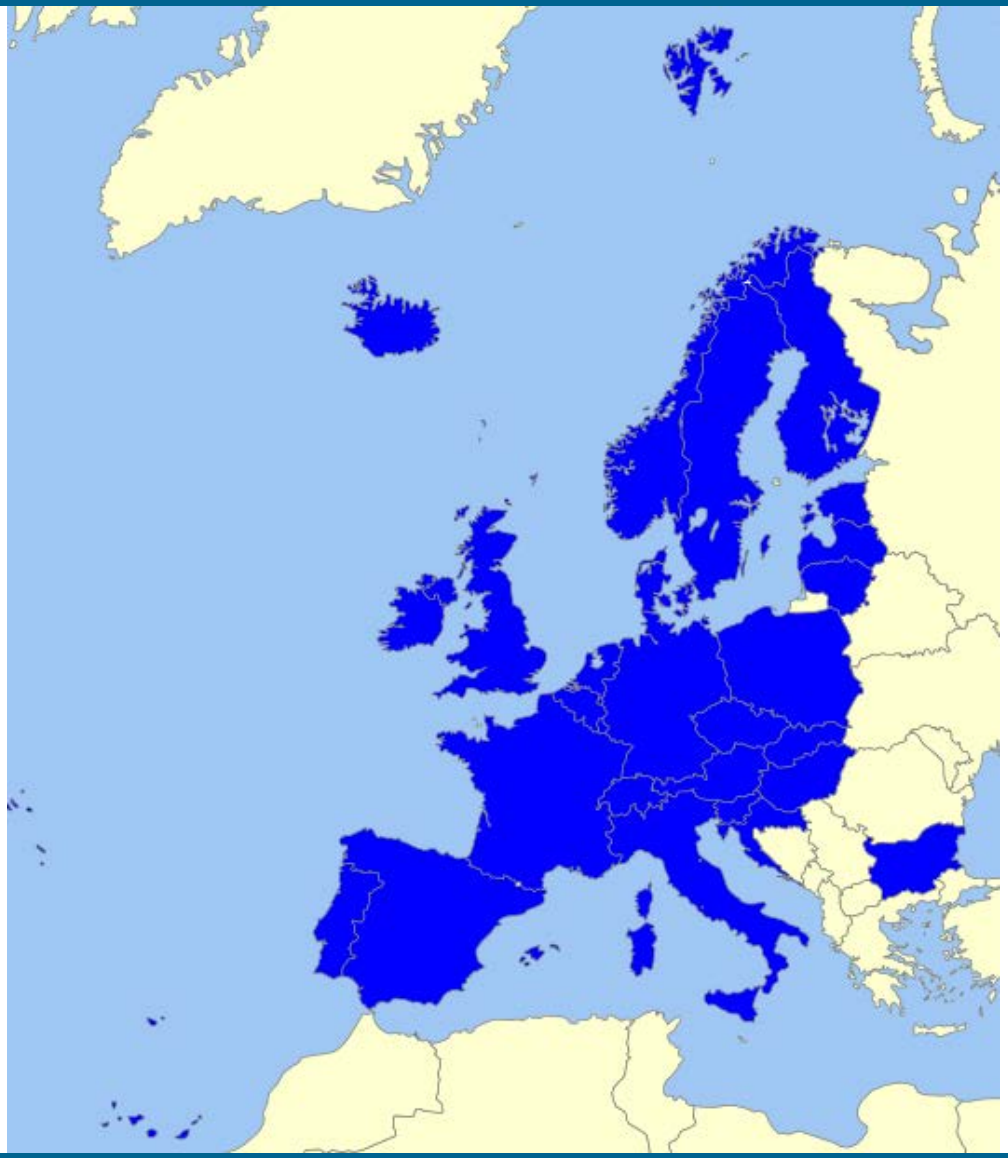
SCIENCE EUROPE
Shaping the future of research

Science Europe in a nutshell

- Association of major public research organisations in Europe: Research Funding Organisations and Research Performing Organisations with a public mission
- 47 member organisations from 27 countries
- Combined R&D investment of appromimately €25 billion per annum
- Policy organisation, no programme management
- Role:
  - Collaboration platform
  - Think-tank for science policy issues
  - Advocacy and interest representation

SCIENCE EUROPE
Shaping the future of research

Members of Science Europe: Country Coverage



SCIENCE EUROPE
Shaping the future of research

Members of Science Europe 2016

Austria	FWF
Belgium	FWO, FNRS
Bulgaria	BAS
Croatia	HRZZ
Czech Republic	GAČR
Denmark	DFF/DCIR, DG
Estonia	ETAG
Finland	AKA
France	ANR, CNRS, CEA, INSERM, IRD
Germany	DFG, MPG, WGL
Hungary	MTA,
Iceland	Rannís
Ireland	HRB, Irish Research Council, SFI
Italy	CNR, INFN
Latvia	LZP
Lithuania	LMT
Luxembourg	FNR
Netherlands	NWO
Norway	RCN
Poland	NCN
Portugal	FCT
Slovakia	APVV, SAV
Slovenia	ARSS
Spain	CSIC
Sweden	FORMAS, FORTE, VR
Switzerland	SNSF
United Kingdom	AHRC, BBSRC, EPSRC, ESRC, MRC, NERC, STFC

SCIENCE EUROPE
Shaping the future of research

Science Europe statement of core principles

- **FAIRness**
- **Flexibility**
- **Data entry minimisation**
- **Openness**

FAIRness

To be Findable:

- common definitions
- F1. (meta)data are assigned a ***globally unique and persistent identifier***
- F2. data are described with rich metadata (defined by R1 in the table)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- forerunner of interoperability
- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- foundations of interoperability
- I1. (meta)data use a formal, accessible, ***shared***, and broadly applicable ***language*** for knowledge representation
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data
- I4. (meta)data can be exchanged through a standard format

To be Reusable:

- characteristics of robust data
- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

Wilkinson, M. D. et al. (2016), *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data 3, 15 March 2016, doi:10.1038/sdata.2016.18

SCIENCE EUROPE

Flexibility

research information systems should be flexible enough to allow for extensions in terms of the data objects covered, their definitions, metadata, and use of external data sources

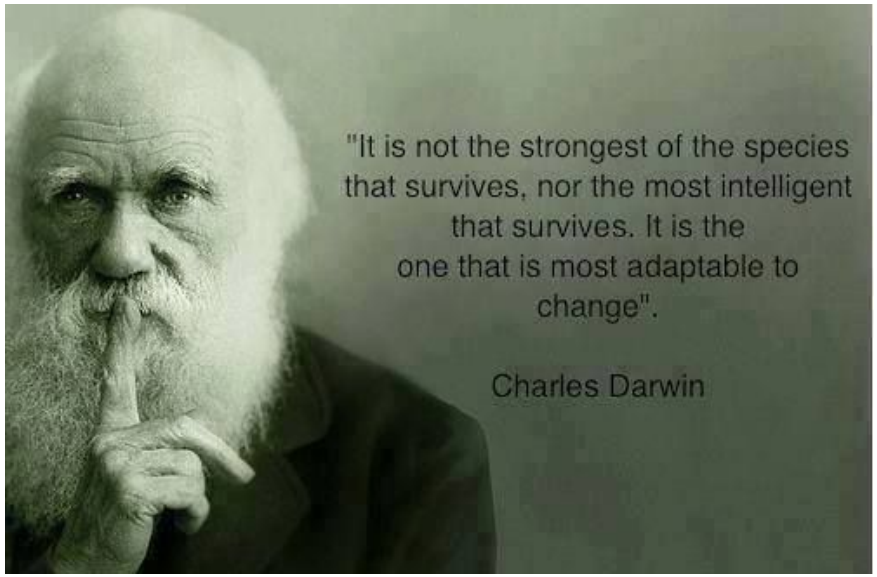
When considering evaluation and analysis

of research funding programmes,

of strategic initiatives,

or of research projects

- needs change
- objectives evolve
- focus shifts



Flexibility

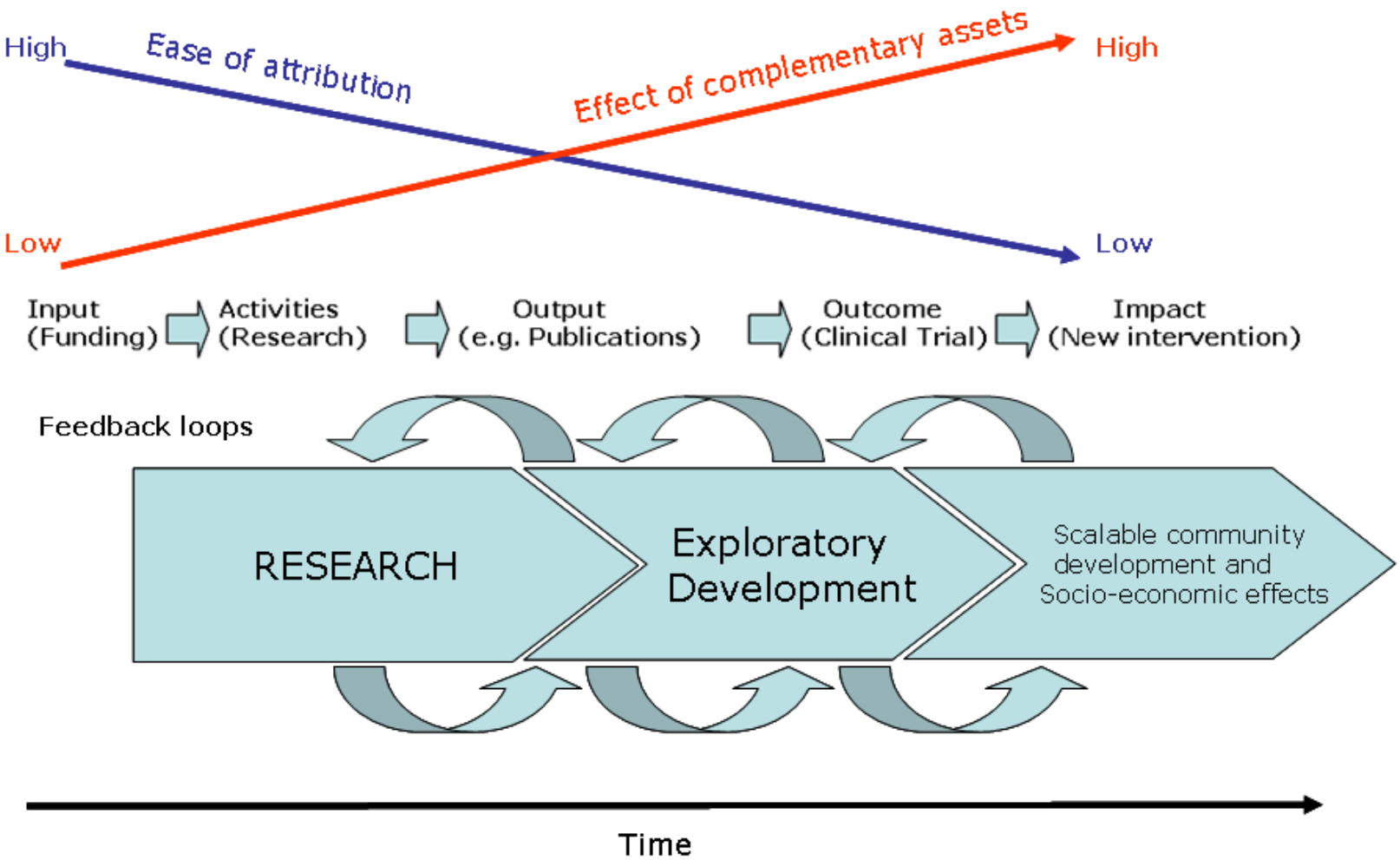
research information systems should be flexible enough to allow for extensions in terms of the data objects covered, their definitions, metadata, and use of external data sources

data + time

reveals impact

Micro-economic approaches

Time, Attribution and Impact (Alan Hughes, 2011)



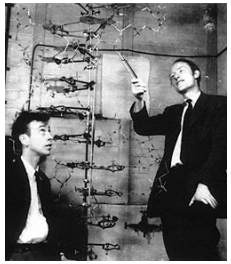
Real-time genetic analysis of the Zika virus

Using the Oxford Nanopore MinION device in front of the minibus lab in Joao Pessoa, Brazil (photo by Ricardo Funari)

- Oxford Nanopore has launched a DNA sequencer the size of a USB stick
- Many potential applications in environmental monitoring, rapid field diagnosis and personalised medicine
- Technology was used in the Ebola epidemic and has been tested in the International Space Station
- In 2015 the MRC funded a project to collect Zika virus samples and genotype them in Brazil - the project's chosen technology was the Oxford Nanopore sequencer

DNA Technologies

UK academic research has created a multi-billion dollar market



1953

Watson and Crick working in MRC laboratories famously describe the structure of DNA

1970

DNA Sequencing

MRC scientist Fred Sanger invents the first viable DNA sequencing technique

The Southern Blot

MRC scientist Ed Southern invents a technique to analyse DNA fragments that still underpins molecular biology today

1980

1984 **DNA fingerprinting** technology invented by MRC-funded scientist Sir Alec Jeffreys at the University of Leicester. Led to research into genetic markers of human diseases

DNA Fingerprinting has revolutionised forensic science, matched donors to patients in need of lifesaving organ transplants, and provided an infallible method of paternity testing

1990

1995 **Oxford Gene Technologies** (OGT) established to commercialise Ed Southern's microarray technology

1998 **Nematode Genome Sequence** UK and US researchers including the MRC Laboratory of Molecular Biology's Dr Sydney Brenner and Sir John Sulston finished sequencing the *C. elegans* genome – the first complete sequence of a multicellular organism lays the foundation for the human genome project

2000

2003 the **Human Genome Project** completes the human genome sequence

2005 **Oxford Nanopore Technologies** (ONT) a spin out based on MRC and EPSRC funded research is bringing to market a transformative next generation sequencing approach

2007 **Solexa** a Cambridge spin out based on BBSRC funded work on SBS sequencing is acquired by US company Illumina

Imperial College spin out DNAe licenses technology to Life Technologies – they develop and market the sequencing approach as Ion Torrent

2010

2005, the UK Forensic Science Service used DNA fingerprinting to match **40,000** samples from crime scenes to people's records on the National DNA Database, including 165 homicides, 100 attempted murders and 570 rapes.

Worldwide market based on genomics and gene-based services is estimated at **£3.5bn**

2012 OGT employs **60 staff** near Oxford and opens a New York office

2011 An economic analysis suggests that the US $3bn investment in the HGP has provided almost **$800bn** of benefit to the US alone

2013 ONT employs **90 staff** in Oxford and Cambridge, and has raised **£105m** since formation in 2005

2012 Roche launches an unsuccessful hostile bid for Illumina of **$6bn**

2013 Life Technologies is acquired by Thermo Fisher for **$14bn**

MRC Medical Research Council

Data entry minimisation

research information systems should **minimise the need for entering data** and facilitate the reuse of data entered manually, in line with the motto '**enter once, reuse multiple times**'.

lessen the burden

to increase quality and compliance

maximise efficient use of resource

Motivation for the development of interoperability

SCIENCE EUROPE
Shaping the future of research

Openness

research information systems' data should be available for external use – in line with the principle 'as open as possible, as closed as necessary' – and their processing should never require the loss of ownership in underlying raw data by the originating institution

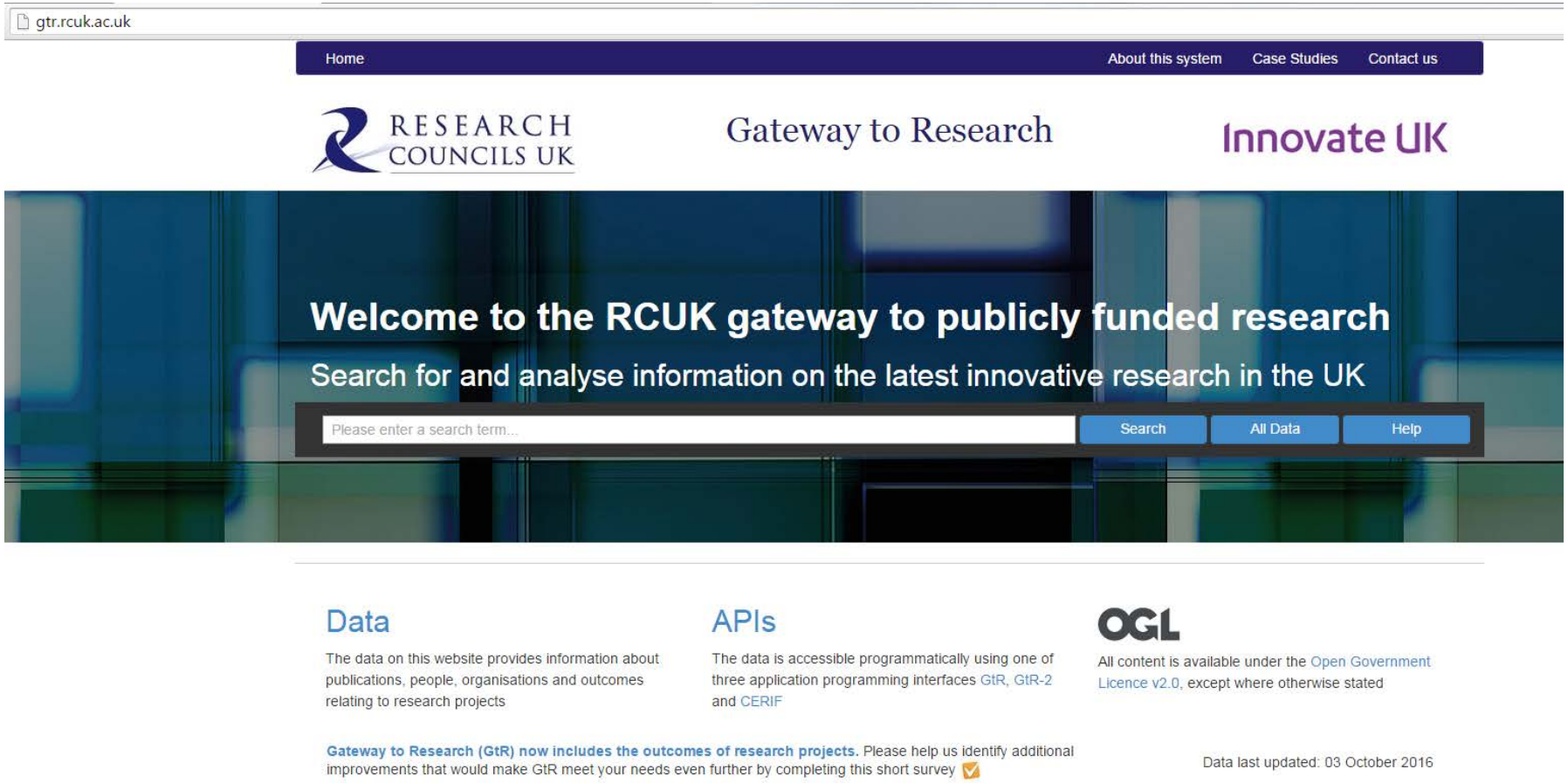
RCUK Gateway to Research

provides a publically accessible portal for RCUK research information

http://gtr.rcuk.ac.uk/

Data sharing

Research Information Management



Searchable for: researcher, funder, keywords in abstracts, research institution, funding amount, award year, award type, etc.

gtr.rcuk.ac.uk

RESEARCH COUNCILS UK

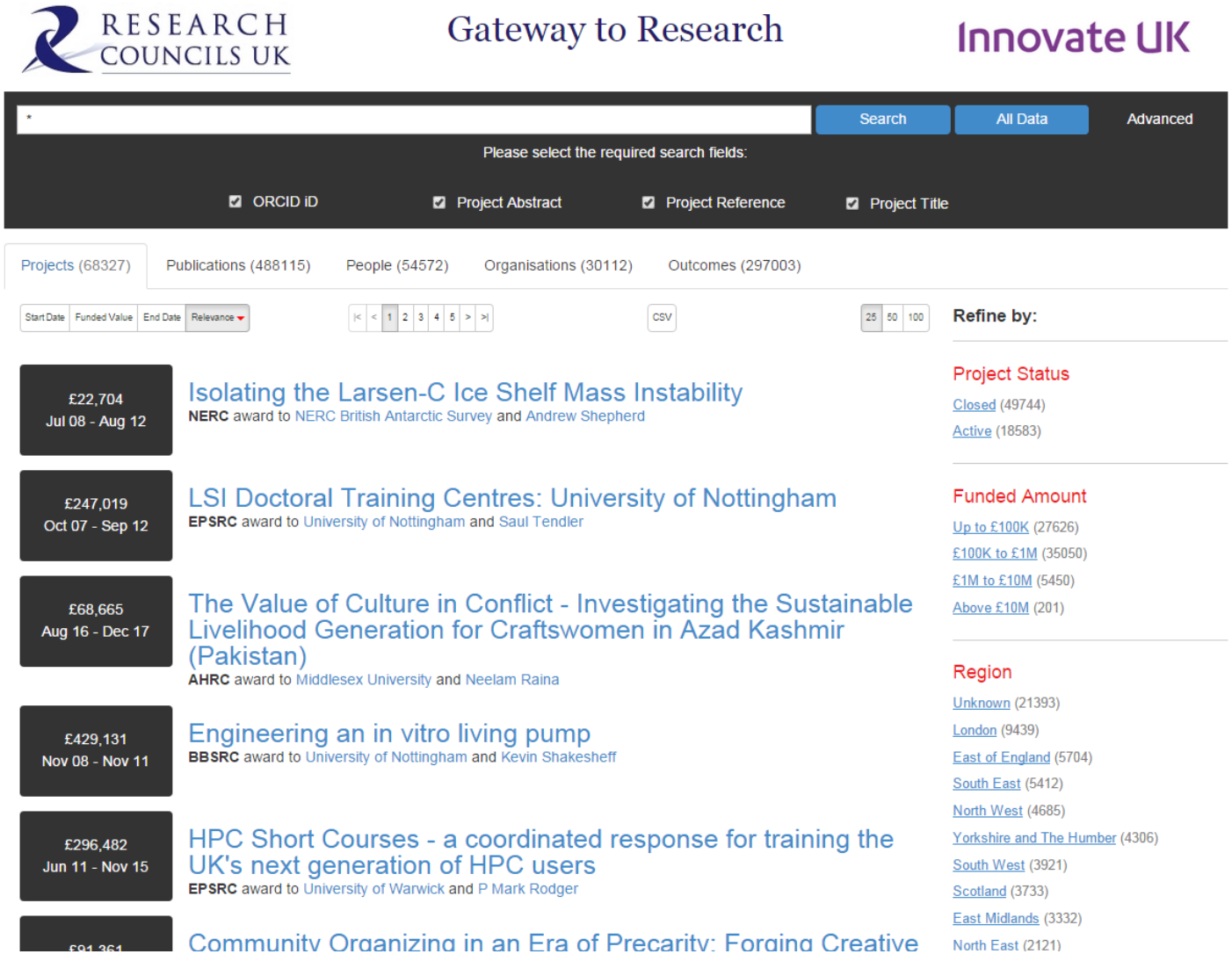
GtR All Data

Research Information Management

as of 3/11/16:
68,000 awards
488,000 publications
54,000 researchers
30,000 RO's
300,000 outcomes

~10,000 visits per month to homepage

reliable API access



laying out the foundations

MRC Medical Research Council

fundamental data management themes

Combining Data

common definitions

publically accessible

interoperability

no pay wall

Funder and Grant Identification

standardised acknowledgement text

CrossRef/DataCite

global grant ID format

Researcher Identification

ORCiD
Connecting Research and Researchers

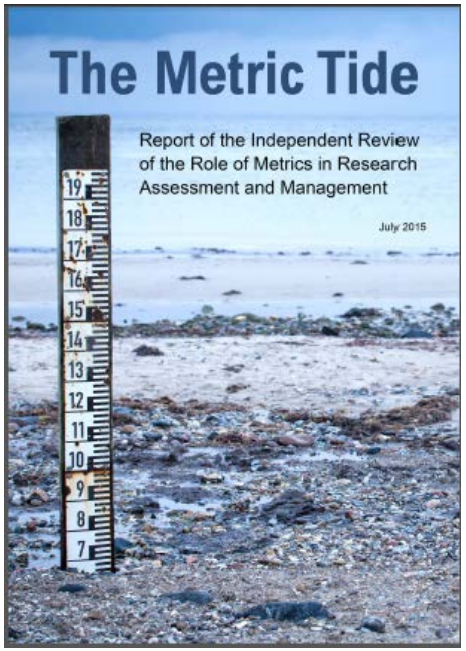
Subject Classifications

comprehensive documentation

cross-mapping

SCIENCE EUROPE

Evaluation context



Indicators can only meet their potential if they are underpinned by an open and interoperable data infrastructure. How underlying data are collected and processed – and the extent to which they remain open to interrogation – is crucial. Without the right identifiers, standards and semantics, we risk developing metrics that are not contextually robust or properly understood. Unique identifiers for individuals and research works will gradually improve the robustness of metrics and reduce administrative burden.

Snowball Metrics

GLOBAL STANDARDS FOR INSTITUTIONAL BENCHMARKING

SCIENCE EUROPE

Interoperability

Interoperability pilot programme 2015

sharing data between

researchfish

RCUK's data collection
and management platform

AND

the distinct research
data repositories of 6
UK Universities

Participants

researchfish

Imperial College London

University of BRISTOL

UNIVERSITY OF CAMBRIDGE

THE UNIVERSITY of EDINBURGH

University of Glasgow

University of St Andrews

The Academy of Medical Sciences

Arts & Humanities Research Council

BBSRC bioscience for the future

CANCER RESEARCH UK

EPSRC Engineering and Physical Sciences Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MRC Medical Research Council

NC 3Rs National Centre for the Replacement Refinement & Reduction of Animals in Research

NERC SCIENCE OF THE ENVIRONMENT

Science & Technology Facilities Council

Summary

- Systematically collecting output data in a structured way has significantly strengthened MRC evaluation capability; researchfish® has proven to be a great success.

- It takes time to build up a useful dataset, and skilled resources to analyse it. Clean and complete data is essential and a good grasp of portfolio/input information.

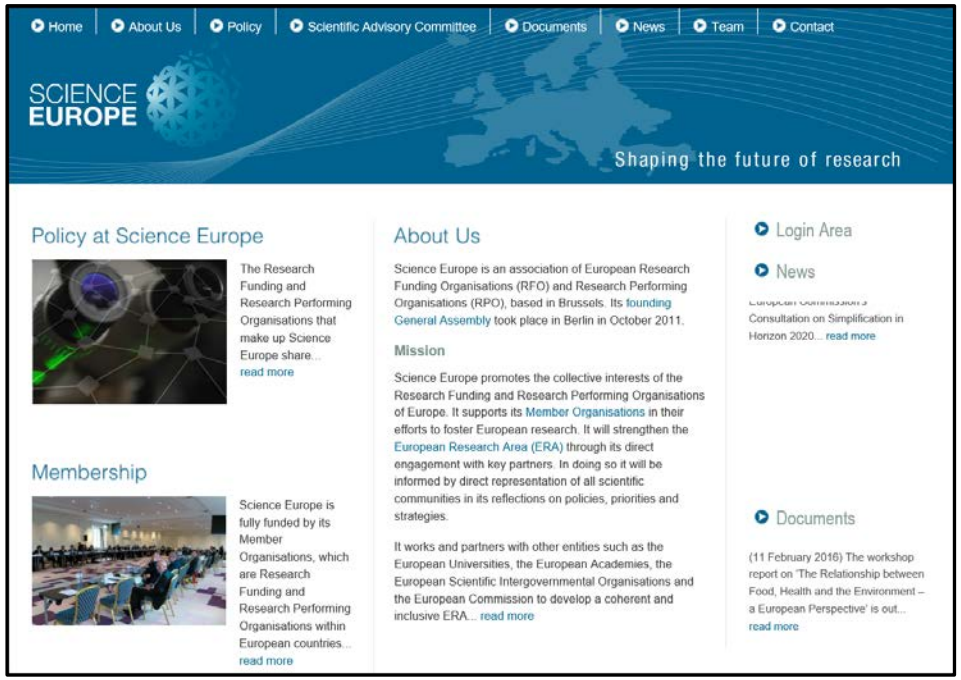
- Outputs other than publications and patents have been found to be the most useful in demonstrating impact, compiling case studies etc.

- Sharing information between funding agencies and universities internationally will be important for understanding what leads to impact, and getting the most out of the data

- More research is needed to assist with micro, meso, and macro-economic views of return on investment, and to propose ways of combining these approaches

Further Information

www.scienceeurope.org



Science Europe
Rue de la Science 14
1040 Brussels
Belgium

+ 32 2 226 0300
office@scienceeurope.org

Researchfish™ health research "output types"

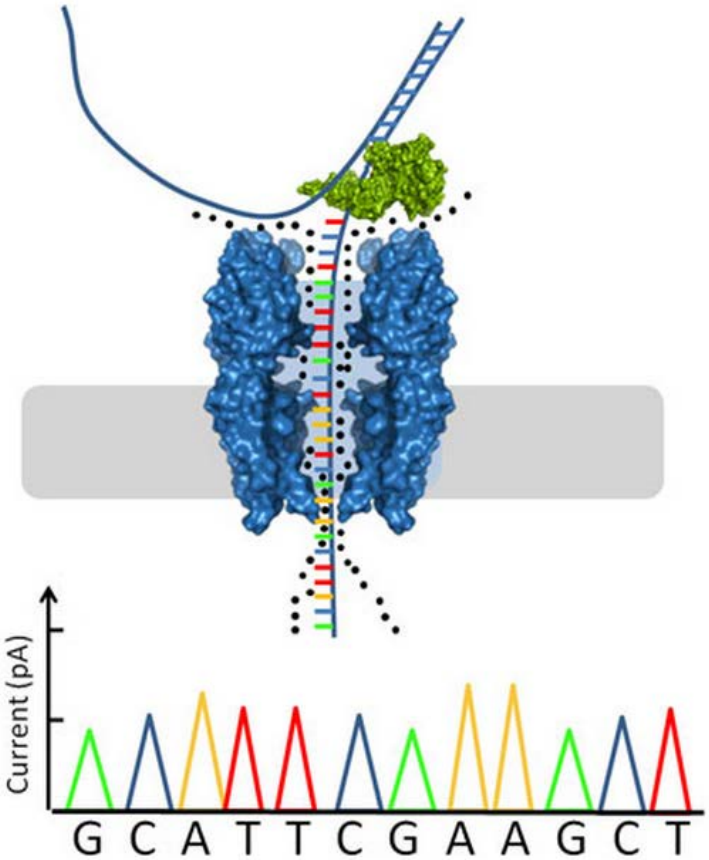
Section	Output type
1	Publications
2	Collaborations
3	Further funding
4	Next destination
5	Engagement activities
6	Influence on Policy, Patients and the Public
7	Research Materials (Tools, Methods, Databases, Models)
8	Intellectual property and licensing
9	Development of products or interventions (medical, artistic, creative, software, clinical trials and technical)
10	Spin Outs
11	Awards and recognition
12	Use of Facilities
13	Other outputs/future steps

MRC research outcomes and impacts

Researchers report progress with, and outcomes/impacts from their MRC-funded research via Researchfish®. Since 2006:

- 60,000 publications, with a FWCI of 2.7 times world average, 60% of these papers are openly accessible in Europe PubMed Central
- 500 patents, approximately 30% of which have been licensed
- 100 spin outs, creating an estimated 1,500 new jobs
- >1100 reported products, interventions and clinical trials in development
  ~150 have reached the market
- >7000 reports of policy influence, 50% of which have impact wider than the UK
- Data has been collected from 6,800 MRC awards, with a total value of £3.7billion. These award have led to >£4.5billion of non-MRC follow on funding (almost 30% from outside the UK)

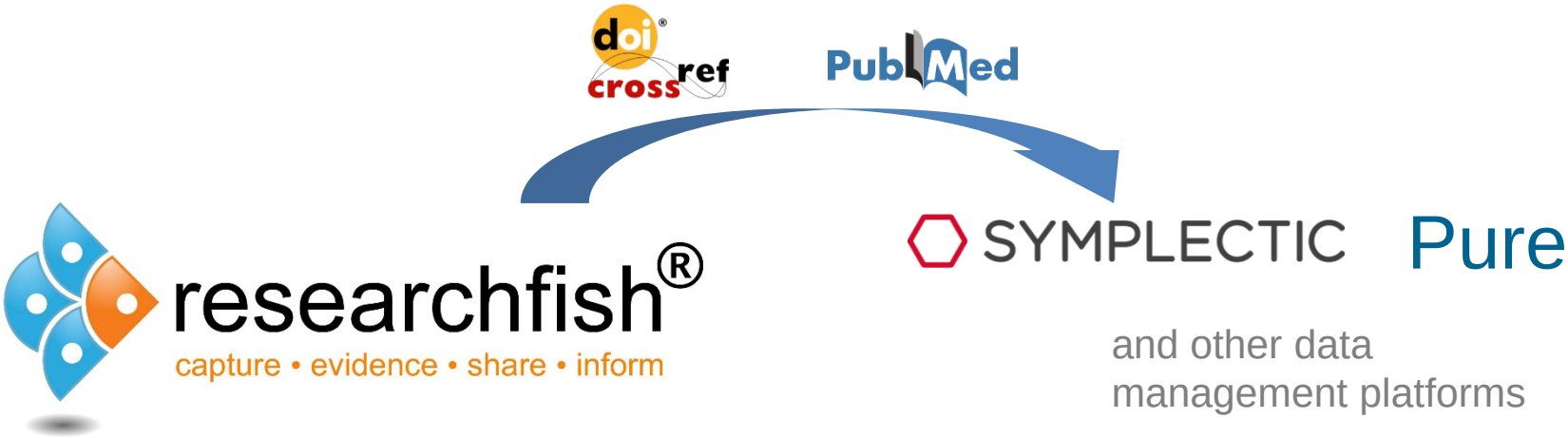
Oxford Nanopore Technologies



A strand of DNA is passed through a nanopore. The current is changed as the bases G, A, T and C pass through the pore in different combinations.

- Chemical biology highlighted as a promising field of interdisciplinary research by MRC/EPSRC prior to 2000
- £1M MRC programme grant to Professor Hagan Bayley at Oxford - use of nanoreactors for studying chemistry at the single molecule level
- One application was a new method of DNA sequencing using nanopore-forming proteins
- In 2005 the University of Oxford formed spin out Oxford Nanopore Technology and by 2010 the company was funding positions in Professor Bayley's laboratory
- By 2016 the company has raised more than **£100M** investment and employs 280 staff

Back-Population of Symplectic

(and other data management platforms)

- “Enter information once and re-use widely” a key tenet
- Existing links within Researchfish back-populated to Symplectic Elements to **existing** awards and publications
- Clear starting point, avoid confusion, demonstrate intent
- Approximately three weeks of work