Referencing software to make it sustainable:

what, why and how?

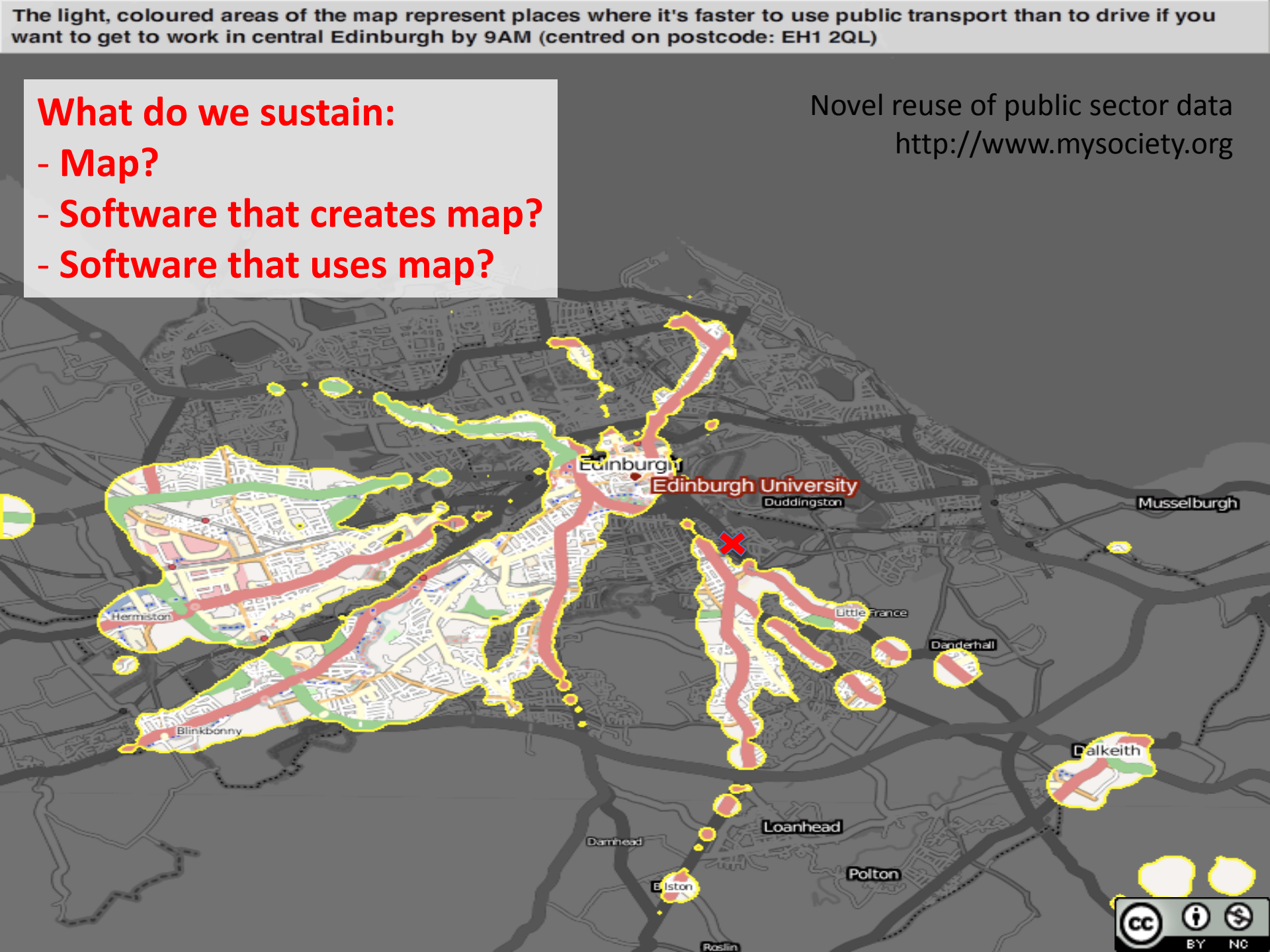
SciencePAD Persistent Identifiers workshop
30 January 2013, CERN

Neil Chue Hong (@npch)
N.ChueHong@software.ac.uk

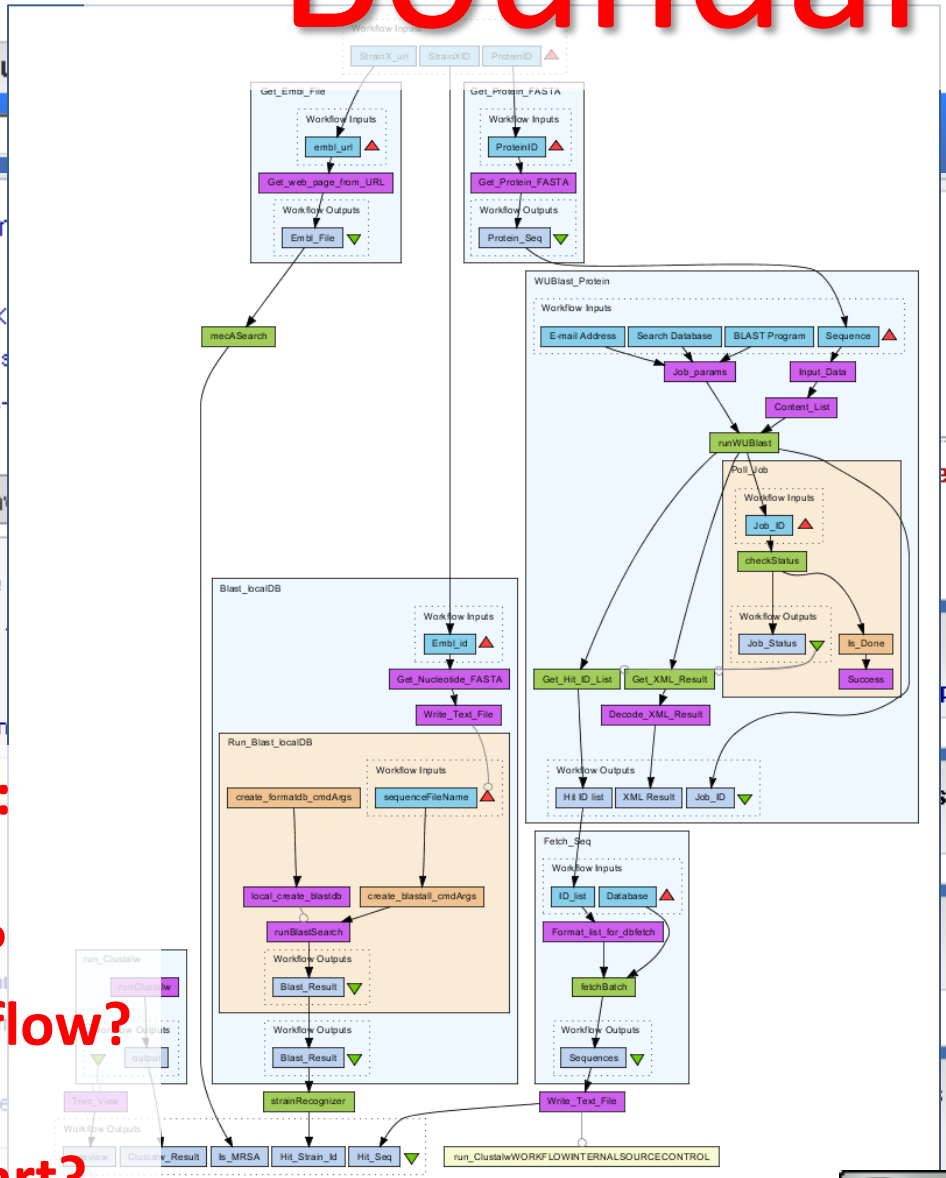
Why is this important?
An animated explanation
Featuring a frustrated panda

http://tinyurl.com/datasharingpanda

Software is no longer easy to define, let alone sustain


The light, coloured areas of the map represent places where it's faster to use public transport than to drive if you want to get to work in central Edinburgh by 9AM (centred on postcode: EH1 2QL)
What do we sustain:
- Map?
- Software that creates map?
- Software that uses map?
Novel reuse of public sector data
http://www.mysociety.org
Edinburgh
Edinburgh University
Duddingston
Musselburgh
Hermiston
Little France
Danderhall
Blinkbonny
Dalkeith
Loanhead
Damhead
Polton
Roslin
BY
NC

Boundary



What do we choose to identify:

- **Workflow?**
- **Software that runs workflow?**
- **Software referenced by workflow?**
- **Software dependencies?**

What's the minimum citable part?

Granularity

Program

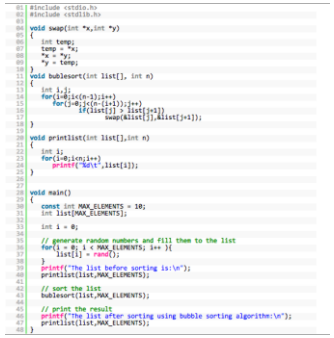
Function

Algorithm

Library / Suite / Package

```c
#include <stdio.h>
#include <stdlib.h>

void swap(int *x,int *y)
{
   int temp;
   temp = *x;
   *x = *y;
   *y = temp;
}

int choose_pivot(int i,int j )
{
   return((i+j) /2);
}

void quicksort(int list[],int m,int n)
{
   int key,i,j,k;
   if( m < n)
   {
      k = choose_pivot(m,n);
      swap(&list[m],&list[k]);
      key = list[m];
      i = m+1;
      j = n;
      while(i <= j)
      {
         while((i <= n) && (list[i] <= key))
                i++;
         while((j >= m) && (list[j] > key))
                j--;
         if( i < j)
                swap(&list[i],&list[j]);
      }
      // swap two elements
      swap(&list[m],&list[j]);
      // recursively sort the lesser list
      quicksort(list,m,j-1);
      quicksort(list,j+1,n);
   }
}
void printlist(int list[],int n)
{
   int i;
   for(i=0;i<n;i++)
      printf("%d\t",list[i]);
}

void main()
{
   const int MAX_ELEMENTS = 10;
   int list[MAX_ELEMENTS];

   int i = 0;

   // generate random numbers and fill them to the list
   for(i = 0; i < MAX_ELEMENTS; i++ ){
       list[i] = rand();
   }
   printf("The list before sorting is:\n");
   printlist(list,MAX_ELEMENTS);

   // sort the list using quicksort
   quicksort(list,0,MAX_ELEMENTS-1);

   // print the result
   printf("The list after sorting using quicksort algorithm:\n");
   printlist(list,MAX_ELEMENTS);
}
```

...



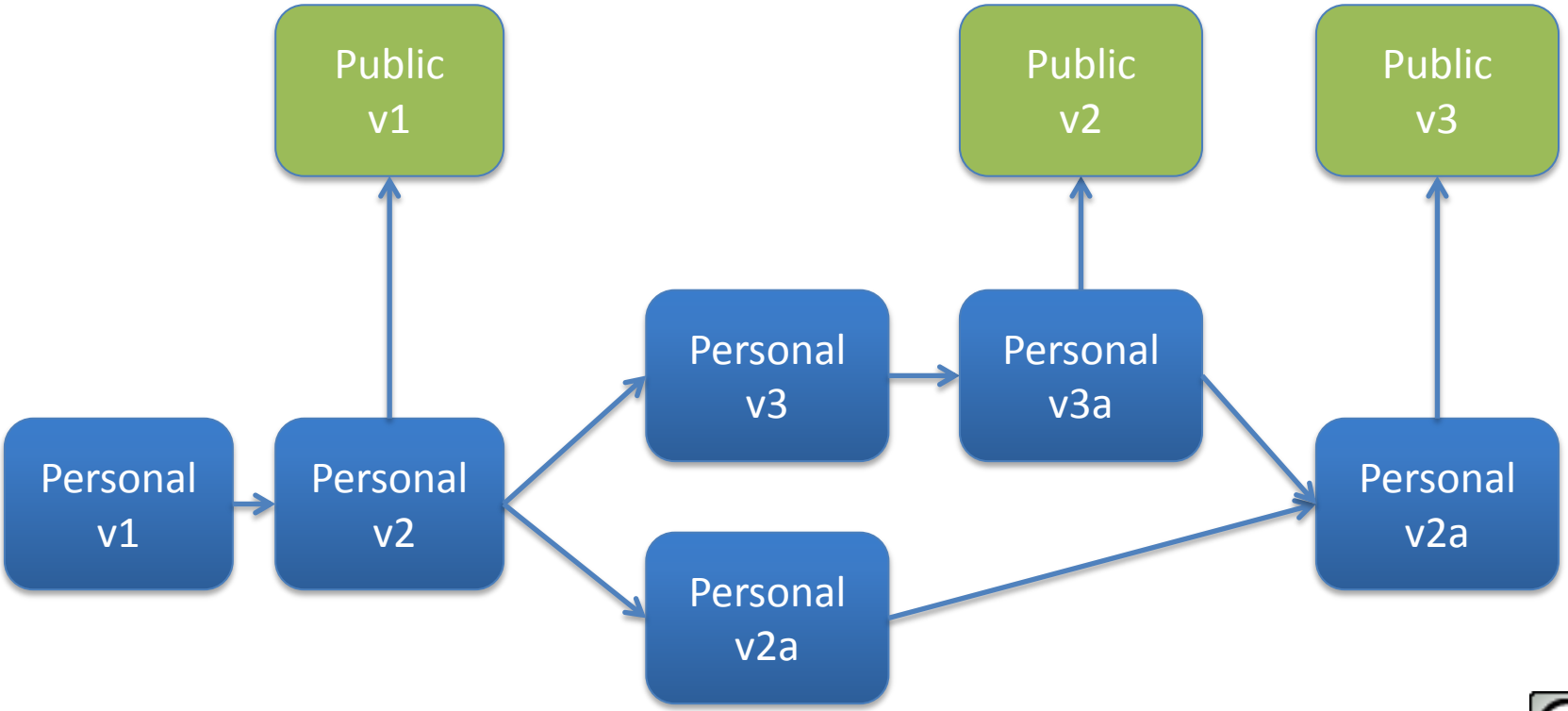
Versioning

Why do we version?

- To indicate a change

- To allow sharing

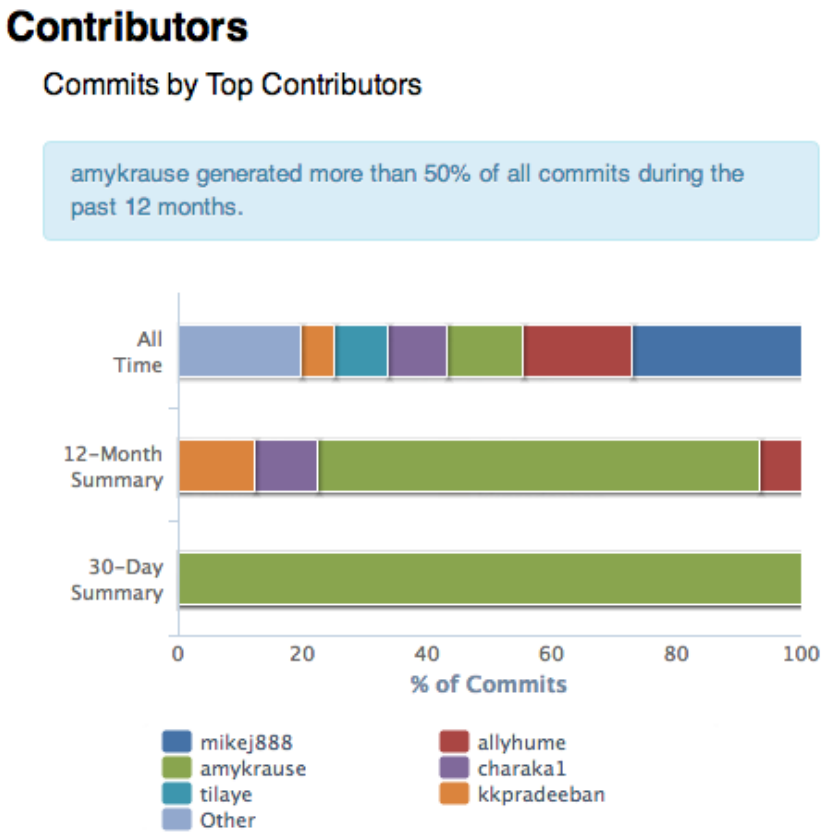
- To confer special status

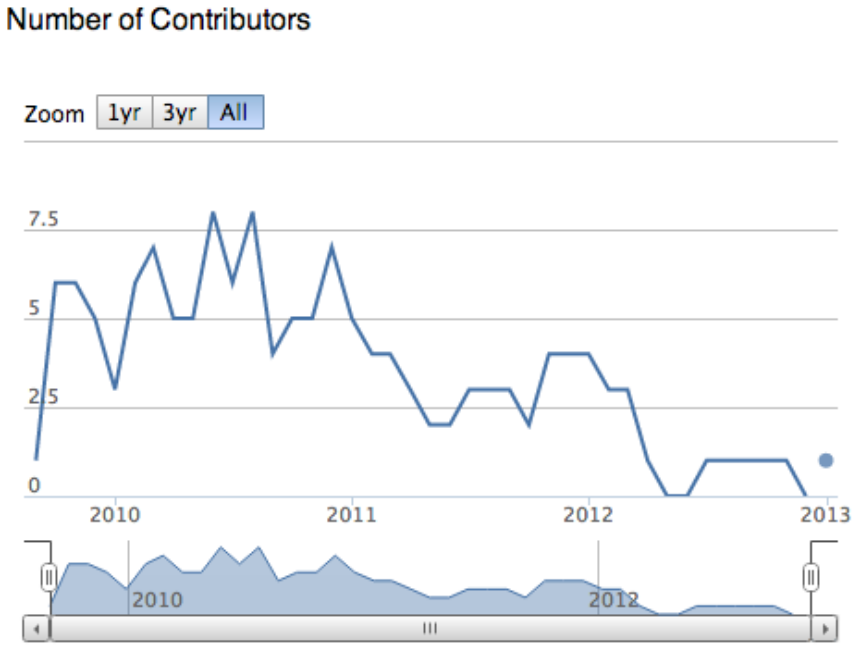


Authorship

- Which authors have had what impact on each version of the software?
- Who had the largest contribution to the scientific results in a paper?

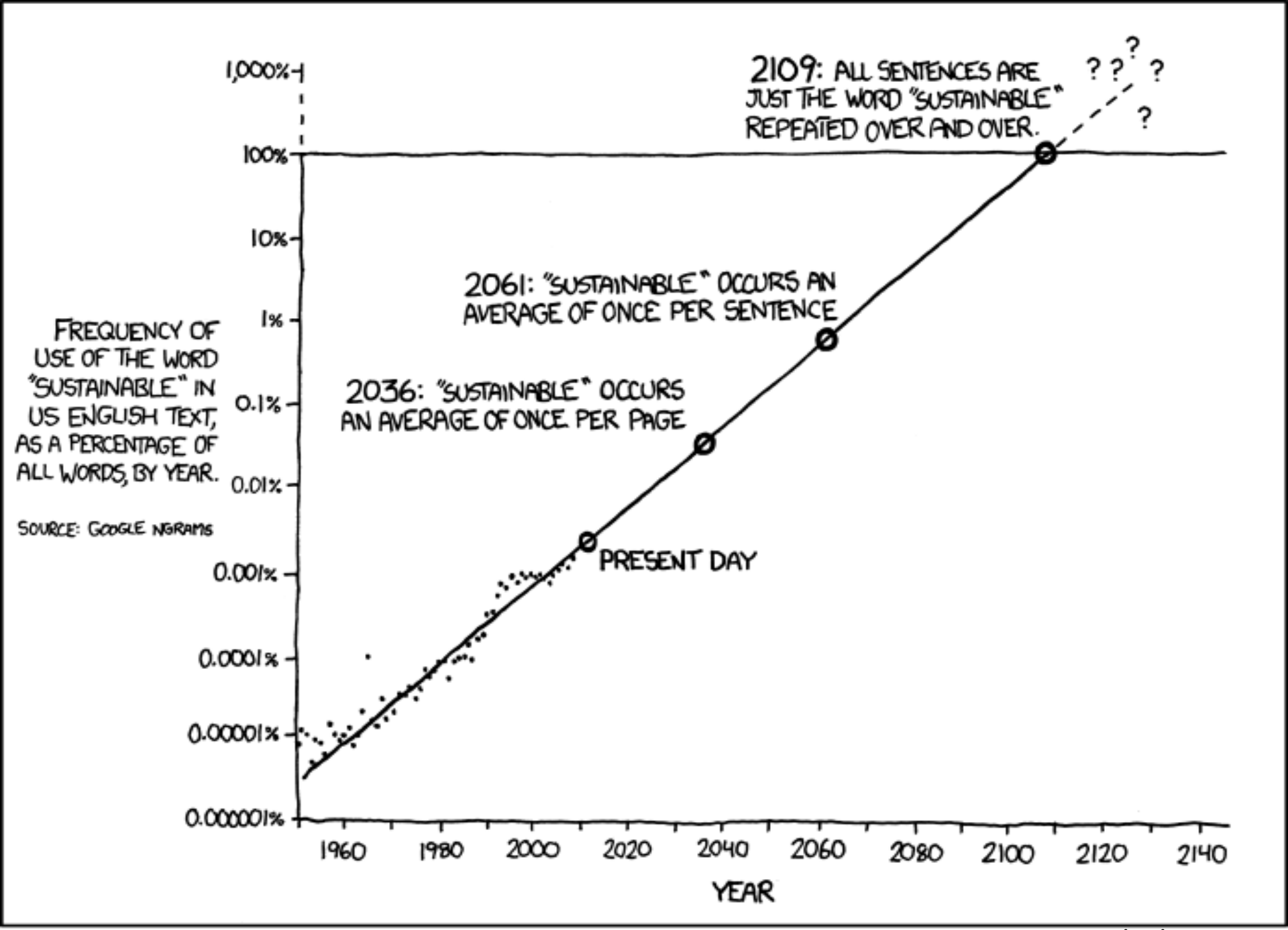
http://beyond-impact.org/?p=175





OGSA-DAI projects statistics from Ohloh

Software sustainability is the ability to continue to use, support, maintain, and evolve software

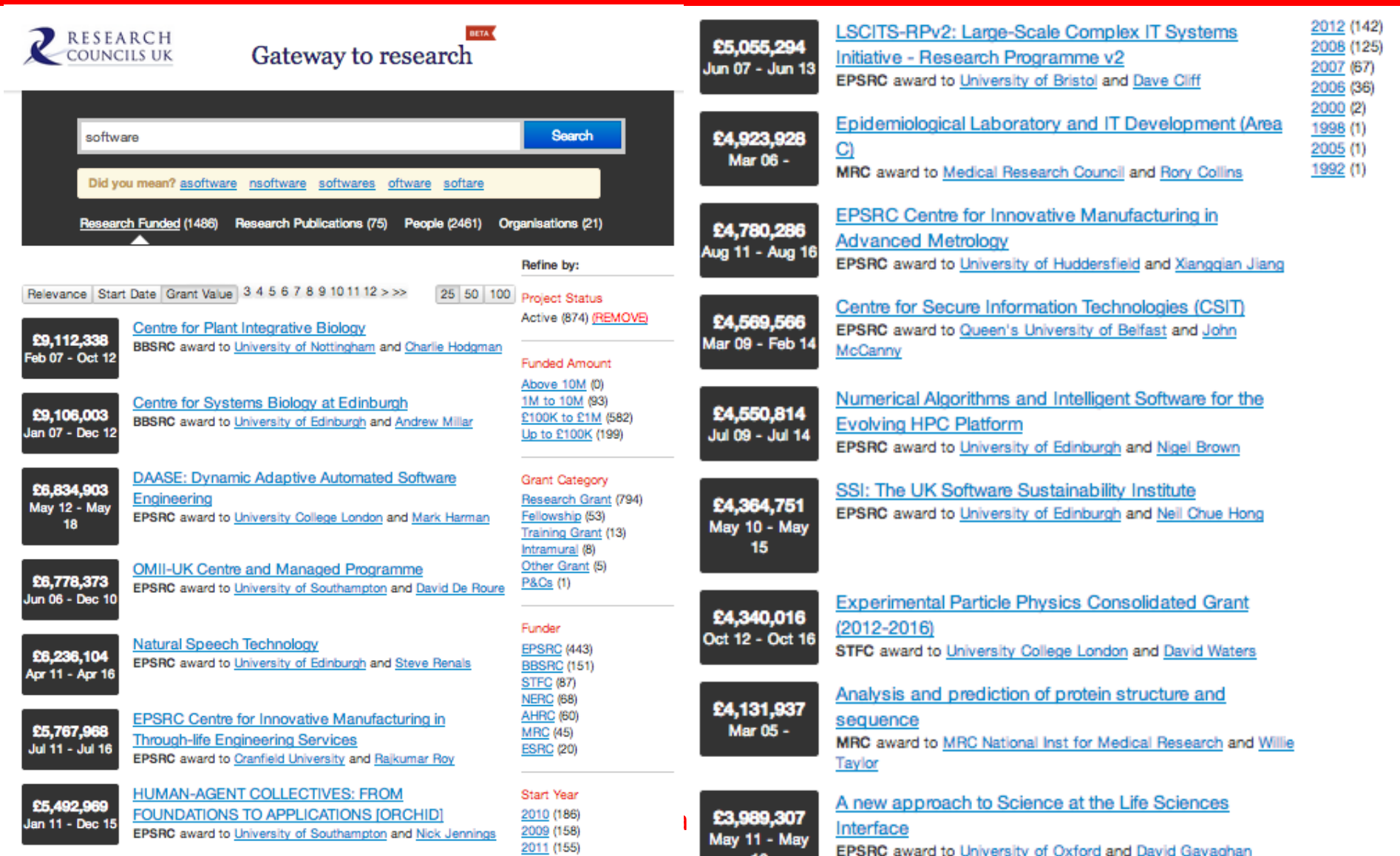


THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

From xkcd.com

RCUK Gateway to Research

www.software.ac.uk



RESEARCH COUNCILS UK

Gateway to research BETA

software

Search

Did you mean? asoftware nsoftware softwares oftware softare

Research Funded (1486) Research Publications (75) People (2461) Organisations (21)

Relevance | Start Date | Grant Value 3 4 5 6 7 8 9 10 11 12 > >> 25 50 100

£9,112,338 Feb 07 - Oct 12 — Centre for Plant Integrative Biology
BBSRC award to University of Nottingham and Charlie Hodgman

£9,106,003 Jan 07 - Dec 12 — Centre for Systems Biology at Edinburgh
BBSRC award to University of Edinburgh and Andrew Millar

£8,834,903 May 12 - May 18 — DAASE: Dynamic Adaptive Automated Software Engineering
EPSRC award to University College London and Mark Harman

£8,778,373 Jun 06 - Dec 10 — OMII-UK Centre and Managed Programme
EPSRC award to University of Southampton and David De Roure

£8,236,104 Apr 11 - Apr 16 — Natural Speech Technology
EPSRC award to University of Edinburgh and Steve Renals

£5,767,968 Jul 11 - Jul 16 — EPSRC Centre for Innovative Manufacturing in Through-life Engineering Services
EPSRC award to Cranfield University and Rajkumar Roy

£5,492,969 Jan 11 - Dec 15 — HUMAN-AGENT COLLECTIVES: FROM FOUNDATIONS TO APPLICATIONS [ORCHID]
EPSRC award to University of Southampton and Nick Jennings

Refine by:

Project Status
Active (874) (REMOVE)

Funded Amount
Above 10M (0)
1M to 10M (93)
£100K to £1M (582)
Up to £100K (199)

Grant Category
Research Grant (794)
Fellowship (53)
Training Grant (13)
Intramural (8)
Other Grant (5)
P&Cs (1)

Funder
EPSRC (443)
BBSRC (151)
STFC (87)
NERC (68)
AHRC (60)
MRC (45)
ESRC (20)

Start Year
2010 (186)
2009 (158)
2011 (155)

£5,055,294 Jun 07 - Jun 13 — LSCITS-RPv2: Large-Scale Complex IT Systems Initiative - Research Programme v2
EPSRC award to University of Bristol and Dave Cliff

£4,923,928 Mar 06 - — Epidemiological Laboratory and IT Development (Area C)
MRC award to Medical Research Council and Rory Collins

£4,780,286 Aug 11 - Aug 16 — EPSRC Centre for Innovative Manufacturing in Advanced Metrology
EPSRC award to University of Huddersfield and Xiangqian Jiang

£4,569,566 Mar 09 - Feb 14 — Centre for Secure Information Technologies (CSIT)
EPSRC award to Queen's University of Belfast and John McCanny

£4,550,814 Jul 09 - Jul 14 — Numerical Algorithms and Intelligent Software for the Evolving HPC Platform
EPSRC award to University of Edinburgh and Nigel Brown

£4,364,751 May 10 - May 15 — SSI: The UK Software Sustainability Institute
EPSRC award to University of Edinburgh and Neil Chue Hong

£4,340,016 Oct 12 - Oct 16 — Experimental Particle Physics Consolidated Grant (2012-2016)
STFC award to University College London and David Waters

£4,131,937 Mar 05 - — Analysis and prediction of protein structure and sequence
MRC award to MRC National Inst for Medical Research and Willie Taylor

£3,989,307 May 11 - May — A new approach to Science at the Life Sciences Interface
EPSRC award to University of Oxford and David Gavaghan

2012 (142)
2008 (125)
2007 (67)
2006 (36)
2000 (2)
1998 (1)
2005 (1)
1992 (1)

The Software Sustainability Institute

www.software.ac.uk

A national facility for cultivating world-class research through software

- Better software enables better research
- Software reaches boundaries in its development cycle that prevent improvement, growth and adoption
- Providing the expertise and services needed to negotiate to the next stage
- Developing the policy and tools to support the community developing and using research software

Supported by EPSRC Grant EP/H043160/1

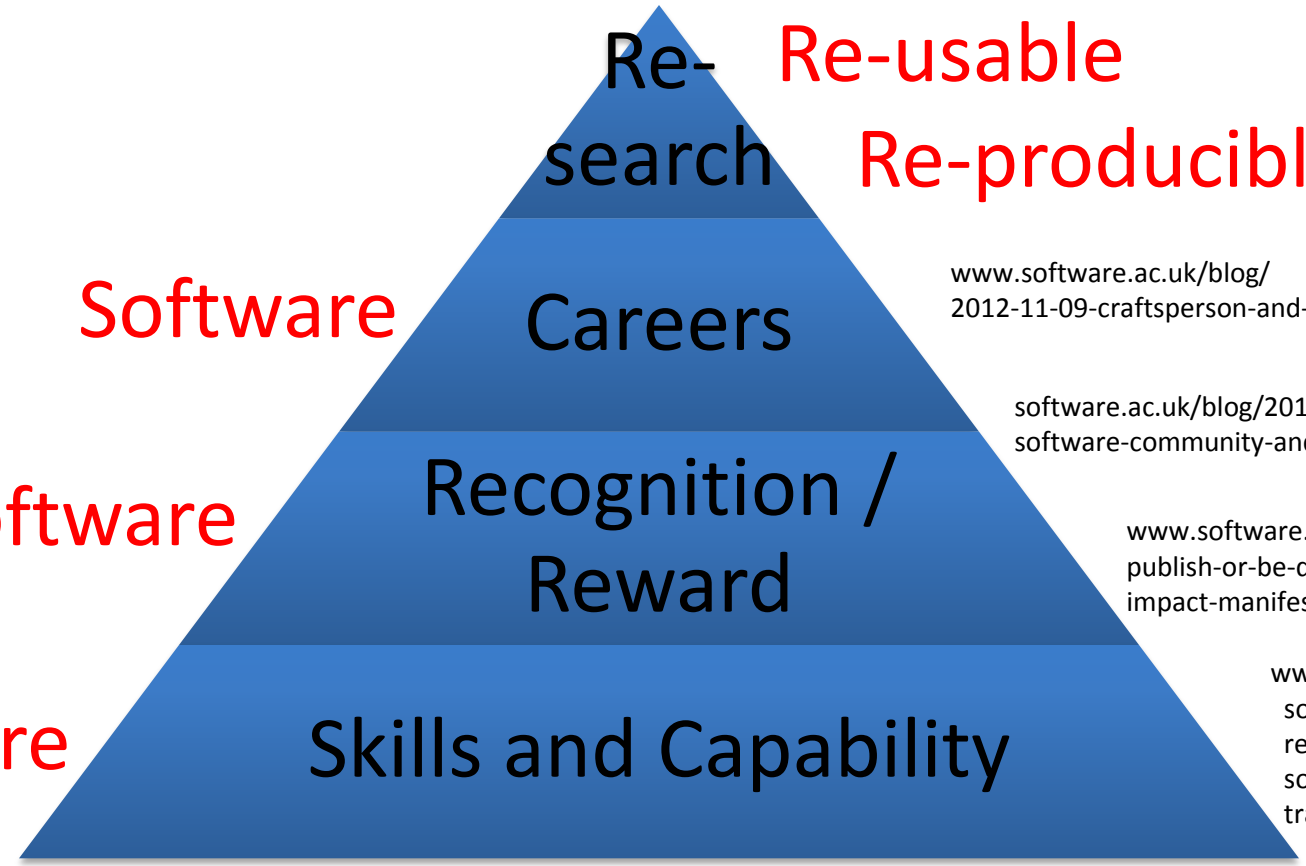
SSI: Long Term Goals

- Provision of useful, effective services for research software community
- Development and sharing of research community intelligence and interactions
- Promotion of research software best practice

- Mantra:
  - Keep the software in its respective community
  - Work with the community, to increase ability
  - Don’t introduce dependency on SSI as the developer
  - Expand and exploit networks and opportunities

The Foundations of Digital Research

www.software.ac.uk



Re-usable

Re-producible

Software

www.software.ac.uk/blog/2012-11-09-craftsperson-and-scholar

software.ac.uk/blog/2012-08-16-what-research-software-community-and-why-should-you-care

Software

www.software.ac.uk/blog/2011-05-02-publish-or-be-damned-alternative-impact-manifesto-research-software

Software

www.software.ac.uk/software-evaluation-guide
resources/guides
software-carpentry
training

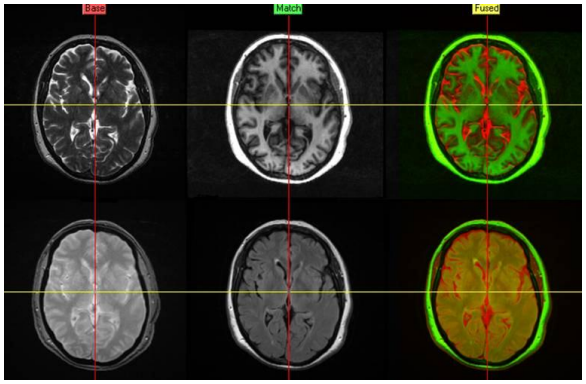
Prlić A, Procter JB (2012) Ten Simple Rules for the Open Development of Scientific Software PLoS Comput Biol 8(12): e1002802. doi:10.1371/journal.pcbi.1002802

Wilson G, et al. (2013) Best Practices for Scientific Computing Submitted to PNAS. http://arxiv.org/abs/1210.0530

Case Study: Brain Imaging

www.software.ac.uk

- Brain Research Imaging Centre, Edinburgh
  - Develop PrivacyGuard software, a DICOM image deidentification toolkit
  - Created software to support new multispectral colouring modulation and variance identification technique ("MCMxxxVI") to identify white matter lesions that are indicative of declining cognitive ability
  - BRIC are not principally software developers, but do provide software to other researchers



clinical research facility EDINBURGH

- SSI's work means the software has been reviewed and refactored
  - Looked at exploitation
    - Usability review, Naming/trademark review
  - Made it easier for BRIC staff to maintain and develop
    - Move to standard repositories, testing and documentation processes
    - Examination of licencing for MCMxxxVI
    - Extraction and refactoring to create standalone tools

- http://www.software.ac.uk/who-do-we-work/brain-research-imaging-centre-edinburgh
- http://www.bric.ed.ac.uk/

Software sustainability requires thriving communities of users and developers: how do we do this?

5 Stars of Scientific Software

www.software.ac.uk

- We need a 5 stars for software!
  - Existence – there is accurate metadata that defines the software
  - Availability – you can access and run the software
  - Openness – the software has an open permissible license
  - Linked – related data, dependencies and papers are referenced
  - Assured – the software provides ways of demonstrating “correctness”

c.f.
5 Stars of Linked Data (Berners-Lee)
5 Stars of Online Journals (Shotton)

Discoverable Software?

www.software.ac.uk

- To grow a community around software, first it must be discoverable
  - For users, wanting to find a solution
  - For developers, wanting to reuse or extend
  - For funders, wanting to promote or feature
- For sustainability
  - Provide useful information
  - Make it easier to attract and add contributors
  - Enable dormant projects to re-activate?

Software Hub Prototyping

www.software.ac.uk

Jisc Software Hub How do people search for software?

Search | Search

Xerte

What information is useful?
Do both provider and user benefit?

Posted January 22, 2013 by admin & filed under Authoring, Learning and Teaching, Open Educational Resources.

Xerte Online Toolkits is a server-based suite of tools for content authors. Elearning materials can be authored quickly and easily using browser-based tools, with no programming required. Xerte Online

Openness Rating

legal	77%
governance	38%
standards	67%
knowledge	72%
market	74%

about the Openess Rating

Programming Factoids

Increasing year-over-year development activity

Large, active development team

Well-established codebase

What can be imported from other sites?
What metadata must be collected to produce this information?

Levels of Showcasing

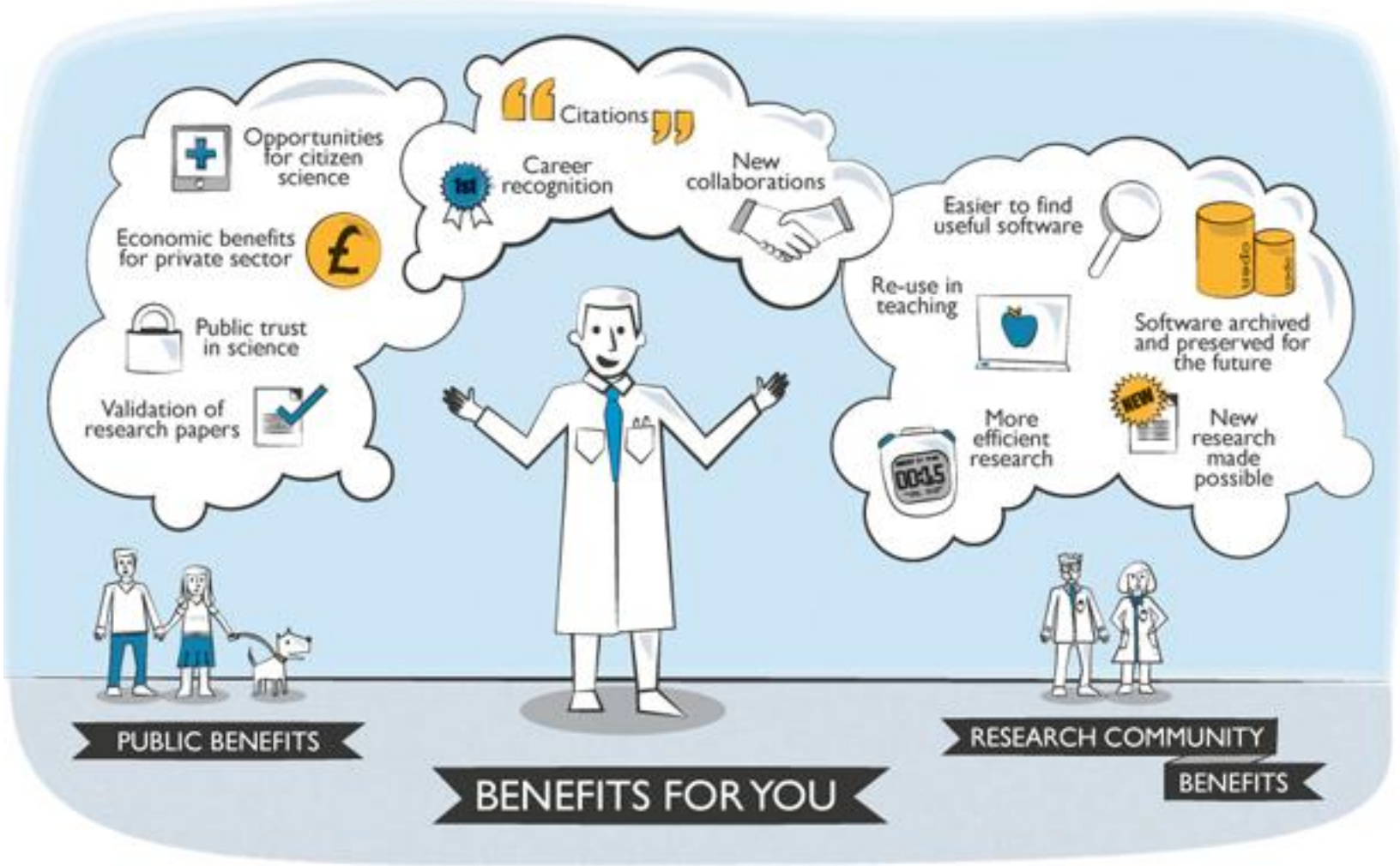
- Level 1: internal
  - Has had support from Jisc
  - Has produced a software output
  - Metadata is incomplete
- Level 2: awaiting approval
  - enough metadata to publish it externally
  - perhaps not all quality criteria met
- Level 3: published
  - meets quality criteria
  - enough information to allow comparison
- Level 4: featured
  - seen as particularly useful, exciting, best of breed etc.
  - associated screencasts, tutorials to show off
- Offer incentives to move up the levels

Collecting Software Metadata

- Can we make the software metadata collection process work?
  - What are the benefits to provider and user?
  - Distinction between Project information and Product Information
  - Difference between information that enables discovery and choice, and the metadata that allows this information to be displayed
    - E.g. “vitality” of project different for developer vs user

Citable Software?

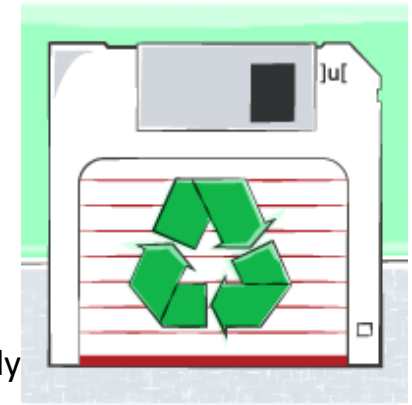
www.software.ac.uk



Software Metapapers

www.software.ac.uk

- Create a complete scholarly record including “standard” publication, method, dataset and models, and software
  - e.g. modelling and simulation, statistical analysis
  - Enable replay, reproduction and reuse
- Pragmatic approach is to create a metadata record for the software, and link it to a copy of the software in some storage infrastructure
  - This is a software metapaper
  - Peer-review the metadata, not the software
- Journal of Open Research Software:
  - http://openresearchsoftware.metajnl.com/



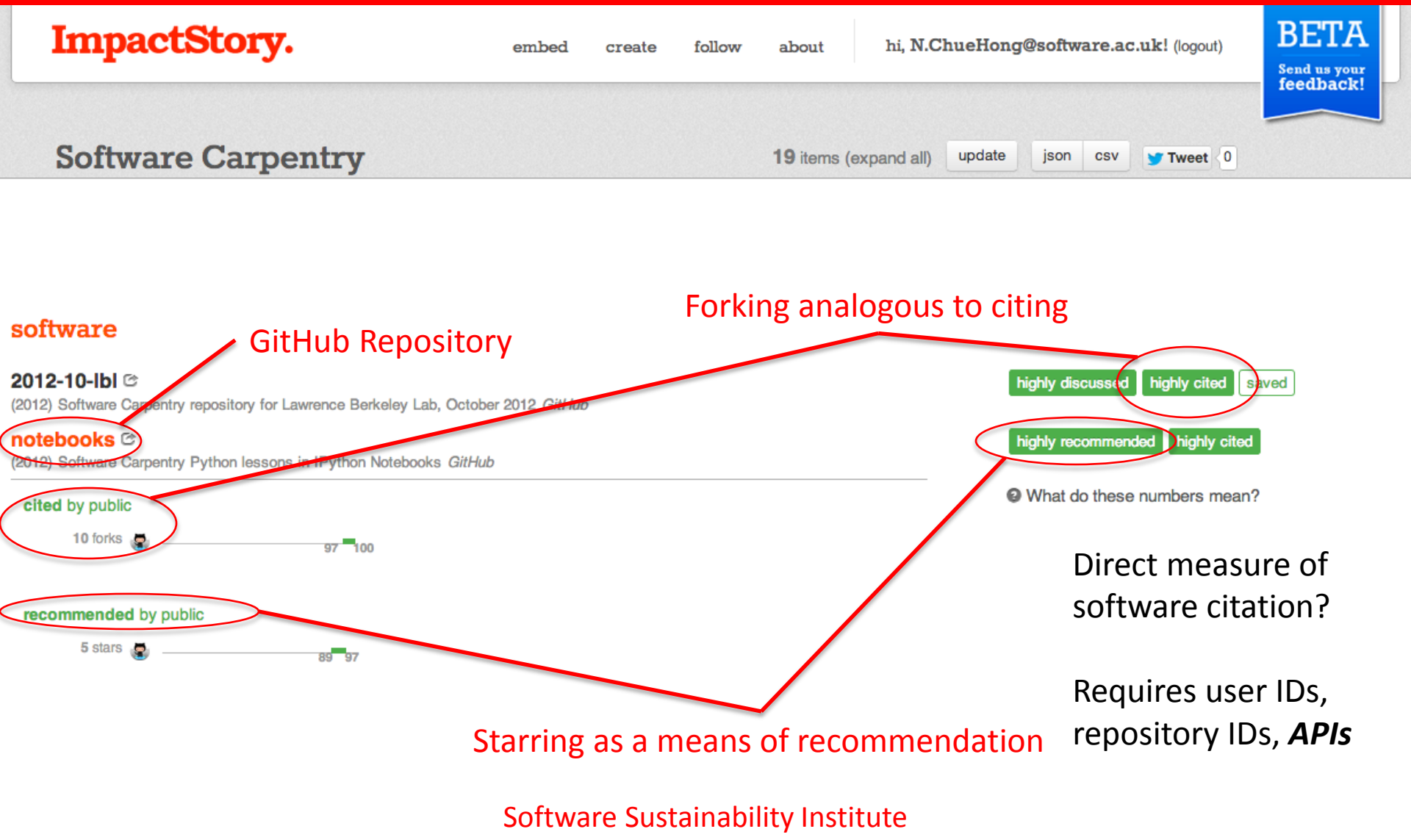
See: http://openresearchsoftware.metajnl.com/faq/
and the work by B. Matthews et al: The Significant Properties of Software: A Study

SoftwareCite

- Does the DataCite approach work with software?
  - What is the cost of minting a DOI?
  - What level do you mint DOIs for software?
  - What is the cost of storing the metadata associated with a software asset?
  - What is the cost of a software asset associated with a DOI disappearing?

Alternative Impact Stories

www.software.ac.uk



Direct measure of software citation?

Requires user IDs, repository IDs, ***APIs***

Software Sustainability Institute

Software sustainability *requires* identifiers (and open APIs) for discovery, growth, attribution + reward

Why is this important?
An animated explanation
Featuring a frustrated panda

http://tinyurl.com/datasharingpanda