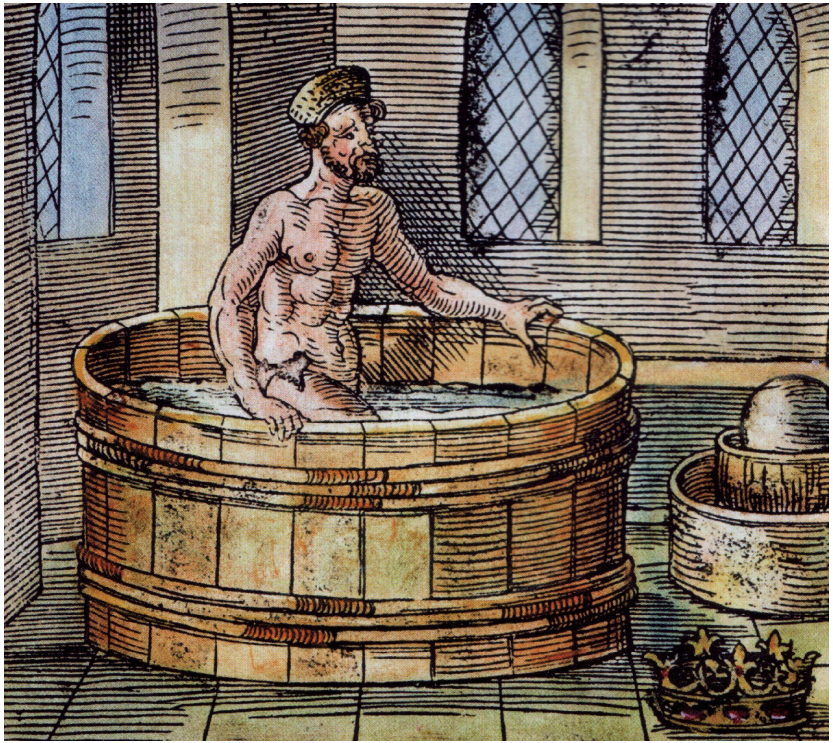
Data Journals y *data papers*: ¿nuevas revistas, viejas costumbres?

Alexandre López-Borrull alopezbo@uoc.edu

CRECS 2015 Murcia, 7 y 8 Mayo de 2015

Ciencia vs Datos

- Relación Nueva? No
- Escala
- Transparencia y eficiencia



http://es.wikipedia.org/wiki/%C2%A1Eureka!#/media/File:Archimede_bain.jpg

Estrategias relación revistas y datos

Pampel & Dallmaier-Tiessen (2014) describen las tres posibilidades por las cuales los científicos pueden difundir sus datos de investigación:

1- Publicación como objeto de información independiente en un repositorio de datos de investigación.

2- Publicación de datos de investigación en forma de *data paper* en una *data journal*.

3- Publicación de datos de investigación junto al artículo, en la forma de *enriched publication*.

¿Qué son los *data papers*?

Chavan y Penev (2013) definen un *data paper* como “una publicación en una revista el propósito de la cual es describir datos en vez de que informar de una investigación o sus conclusiones”.

Así, contendría los datos, sin las hipótesis y/o los argumentos, los resultados o la discusión que se haya llevado a cabo.

Por extensión, una *Data Journal* sería una revista especializada en la publicación de *Data Papers*.

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN

Sequence variants from whole genome sequencing a large group of Icelanders

Daniel F Gudbjartsson, Patrick Sulem, Hannes Helgason, Arnaldur Gylfason, Sigurjon A Gudjonsson, Florian Zink, Asmundur Oddson, Gisli Magnusson, Bjarni V Halldorsson, Eirikur Hjartarson, Gunnar Th. Sigurdsson, Augustine Kong, Agnar Helgason, Gisli Masson, Olafur Th. Magnusson, Unnur Thorsteinsdottir & Kari Stefansson

Affiliations | Contributions | Corresponding authors

Scientific Data **2**, Article number: 150011 (2015) | doi:10.1038/sdata.2015.11

Received 18 July 2014 | Accepted 04 March 2015 | Published online 25 March 2015

PDF | ISA tab | Citation | Reprints | Rights & permissions | Article metrics

Abstract

We have accumulated considerable data on the genetic makeup of the Icelandic population by sequencing the whole genomes of 2,636 Icelanders to depth of at least 10X and by chip genotyping 101,584 more. The sequencing was done with Illumina technology. The median

Background & Summary

Genome-wide association scans were initially based on 300–600 k SNP chip genotyping arrays designed based on the HapMap dataset[1]. The HapMap project focused primarily on common variants (MAF>5%) and methods were subsequently developed to accurately impute 2.5 million HapMap phase 2 SNPs into such chip data[2]. This has led to the discovery of a plethora of associations between common sequence variants and human diseases and traits[3].

Large scale whole genomic sequencing has allowed the detection of rare sequence variants that range in effect from causing diseases to modifying complex disease risk—variants that would recently either not have been observed or could not be tested for association with disease on a sufficiently large scale. Several large sequencing projects are ongoing such as the 1000 Genomes project[4], the Exome sequencing project (ESP)[5,6] and the GoNL project[7].

We have sequenced the whole genomes of 2,636 Icelanders using Illumina technology. The individuals were selected for sequencing based on having a wide range of phenotypes (Tables 1 and 2). The sequencing was done to a mean depth of at least 10X (median 20X), including 909 to a mean depth of at least 30X (Fig. 1). For individuals with an average depth of at least 10X, a coverage of at least 1X was achieved for 2.72 Gb and of 10X or more for 2.70 Gb. For individuals with an average depth of at least 30X, a coverage of at least 30X was achieved for 2.35 Gb. A total of 20 million autosomal SNPs and 1.5 million indels, up to a length of 60 base-pairs (bp), were identified and their genotypes called for all samples simultaneously using the Genome Analysis Toolkit (GATK version 2.3.9, Fig. 2)[8]. We used information about haplotype sharing, taking advantage of the fact that all the sequenced individuals had also been chip-typed and long range phased to improve variant genotyping[9].

Table 1: The 50 most prevalent conditions among the 2,636 sequenced Icelanders.

Methods

These methods are an expanded version of the descriptions contained in Gudbjartsson *et al.*[13]

The Icelandic study population

This study is based on whole-genome sequence data from the whole blood of 2,636 Icelanders participating in various disease projects at deCODE genetics (Tables 1 and 2). In addition, a total of 104,220 Icelanders have been genotyped using Illumina SNP chips.

All participating individuals, or their guardians, gave their informed consent before blood samples were drawn. The family history of participants donating blood was incorporated into the study by including the phenotypes of first and second degree relatives and integrating over their possible genotypes. This integration is performed without the genotypes being kept in storage.

All sample identifiers were encrypted in accordance with the regulations of the Icelandic Data Protection Authority. Approval for these studies was provided by the National Bioethics Committee and the Icelandic Data Protection Authority.

The Icelandic genealogy

The Icelandic genealogical database contains 819,410 individuals back to 740 AD. Of the 471,284 Icelanders recorded to have been born in the 20th century, 91.1% had a recorded father and 93.7% had a recorded mother in the database. Similarly, of the 183,896 Icelanders recorded to have been born in the 19th century, 97.5% had a recorded father and 97.8% had a recorded mother.

The Icelandic genealogy was extract from many sources. Primarily from church books, censuses, the Registers Iceland (http://skra.is), local records of inhabitants and other official documents, but also from other sources such as old manuscripts, letters, annals, books of Althingi, books of judgments, books of family pedigrees, registers of farmers, registers of professional and lists of

UPmetajournals

Journal of open archaeology data

ubiquity press open access

How to cite | CC-BY 3.0 License and copyright information | Peer reviewed | Publication dates

DATA PAPER

Survey on Material Used to Fill Wooden Objects During Conservation

Kate Fulcher [1]

1. PhD student, University College London, London, United Kingdom

Abstract

The data here described are the results of a survey sent to conservators to discover the range of materials being using to fill gaps in wooden objects, and to assess how popular hydroxypropyl cellulose and paper pulp were in regards to this. The survey was created using Survey Monkey, a free online survey tool. The complete dataset is available from UCL Discovery in Excel and CSV format. It is hoped that these survey results can be used by conservators to inform their work on wooden objects and inspire experimental work that will demonstrate the suitability of different materials for the purpose of filling wood.

Keywords: conservation, wood, heritage, Klucel, cellulose, hydroxypropyl cellulose, filler, survey, microballoons, acrylic resin, Paraloid B72.

(1) Overview

Context

The survey was conducted in April/May 2012. The purpose of this survey was to find out what range of materials conservators were using to fill wood, and to assess how popular hydroxypropyl cellulose and paper pulp were in regards to this. It formed part of a larger, experimental, study into the use of cellulose-based materials to fill wooden objects in conservation. Compensating for loss in objects has a long history, and most likely pre-history. The reasons for loss compensation, or "filling", are usually structural or aesthetic. If the object's stability is threatened by the material loss, or the understanding or interpretation of the object is difficult, the gap may be filled; different material properties for the fill may be required for differing circumstances. By filling the loss, the object is more able to fulfil its intended use (at the time of repair), which can vary from being a cooking utensil, to holding religious significance, to

(3) Dataset description

Object name

- Survey responses wood fills.xlsx
- Intro.csv
- Summary.csv
- Q1.csv
- Q2.csv
- Q3.csv
- Q4.csv
- Chart data.csv

Data type

Primary data and processed data.

Format names and versions

Excel and CSV both available (CSV in 7 files).

Creation dates

Original data created May 2012.

Dataset creator

Kate Fulcher

Language

English

License

Creative Commons Attribution 4.0 International License.

Repository location

UCL Discovery: http://discovery.ucl.ac.uk/1416824/.

Publication date

19 May 2014

(4) Reuse potential

The survey was used by the author to inform experimental work that focussed on the suitability of cellulose based materials (specifically hydroxypropyl cellulose and paper pulp) to fill wooden objects for conservation. It is hoped that this work can be built on and expanded to look at other materials, and that this survey will provide the basis for choosing which materials to investigate.

UPmetajournals

Journal of open psychology data

ubiquity press open access

DATA PAPER

Data from the Race Implicit Association Test on the Project Implicit Demo Website

Kaiyuan Xu [1], Brian Nosek [2], Anthony G. Greenwald [1]

1. University of Washington, Seattle, Washington, United States of America
2. University of Virginia and Center for Open Science, Charlottesville, Virginia, United States of America

Abstract

This data archive includes Race Implicit Association Test (IAT) scores of 2,355,303 Internet volunteers who completed educational/demonstration versions of the Race IAT at https://implicit.harvard.edu from 2002 to 2012. Data in this archive can be downloaded for all years, either separately by year or in a single file. Codebooks, indicating the variable labels and value labels, and changes of variables over years, are available for both individual-year data sets and the entire data set. Participation in the (still on-going) Race IAT "study" at the Project Implicit (PI) demonstration site includes completion of the Race IAT along with demographic questions, self-report measures of racial attitude, and various additional measures received by a portion of the participants. These data allow analyses involving changes in responding over time and interrelations among IAT and self-report measures of race attitudes, as well as the association of each of these with demographics. This archive is available at http://osf.io/project/52qxL/.

Dataset

The Data described in this paper is available from the Open Science Framework: https://osf.io/52qxl/ [1].

Keywords: Implicit Association Test, race, Implicit attitude, IAT.

Funding Statement

Project Implicit, University of Washington, Yale University, University of Virginia, and Harvard University provided resources essential to creating and maintaining the archive.

(3) Dataset description

Object name

Datasets and Codebooks from Race IAT 2002–2012

Data type

Primary data (self-report and demographic responses), and processed data (IAT scores or measures computed from self-report and demographic responses).

Format names and versions

The archived data are in SPSS .sav files created by SPSS Windows Version 19. The codebooks are .xlsx files created by MS Office Excel 2010.

Creation dates

Data were collected from 23 December 2002 to 31 December 2012. The data released on 24 October 2013 were processed in September 2013.

Dataset creators

Kaiyuan Xu retrieved the raw data from the PI database at implicit.harvard.edu, computed IAT scores and IAT-related variables, provided variable labels and value labels for all measures, and created codebooks for each year's data sets. Brian Nosek provided files that had archived procedural changes over the years. Tony Greenwald provided IAT scoring syntax for SPSS, and advised on formatting of datasets and codebooks.

Language

English

License

CC-BY

Embargo

None

Repository location

Open Science Framework: https://osf.io/52qxl/ [1].

Publication date

Some of the archived data (relatively small portions of the total) have been partially reported in various prior publications:

- Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). 27 February 2002 [6]
- Nosek, B. A., et al. (2007). 07 April 2008 [7]
- Schmidt, K., & Nosek, B. A. (2010). 18 January2010 [8]

Ventajas *Data Papers*

- Mayor profundidad en aspectos de metodología
- Experimentos que no dieron resultado, tener los datasets para comparaciones
- Favorece la reutilización
- Existe revisión de los artículos
- Permite la difusión de los *datasets* con valor añadido
- No sustituye al *full paper*
- Sinergia con los repositorios de datos, tanto generales como especializados
- Emergencia de datos antiguos no publicados

Riesgos *Data Papers*

- Campo abonado a *predatory journals*
- Riesgo de abonar “salami publication”
- ¿Cómo se revisan los paquetes?
- ¿En qué formatos son disponibles?
- ¿Quién se responsabiliza de los aspectos éticos y de privacidad de los *datasets* (revista, autor…)?
- Enlace y vínculo con el *full paper*

Conclusiones

- Una nueva especie en el ecosistema de la Comunicación Científica
- Se necesita perspectiva para poder evaluar la implantación
- Estudios de calidad y diferenciación entre iniciativas

Referencias

Chavan, V.; Penev, L. (2011). “The Data Paper: a mechanism to incentivize data publishing in biodiversity science”. *BMC Bioinformatics*, num. 12 (Suppl15, S2).

Newman, P.; Corke, P. (2009) “Editorial: Data Papers - Peer Reviews Publication of High Quality Data Sets”. *The International Journal of Robotics Research*, vol. 28, num. 5, p.587.

Pampel, H; Dallmaier-Tiessen, S. (2014) “Open Research Data: From Vision to Practice”. En: Bartling, S.; Friesike, S (eds.). *Opening Science* (2014). http://www.

Muchas gracias!

@alexandrelopez