IST AUSTRIA
Institute of Science and Technology

**Document Name** Actual_state_of_research_data_@_IST_Austria

**Author** Jana Porsche
**Version** 1.0
**Modified** 12.11.2012
**Location** http://repository.ist.ac.at/id/eprint/103

**Table of Contents**

# Purpose

This document is created as a part of the project “Repository for Research Data on IST Austria”. It summarises the actual state of research data at IST Austria, based on survey results. It supports the choice of appropriate software, which would best fit the requirements of their users, the researchers.

Currently almost 50% of research groups at IST Austria (11 out of 24) are receiving ERC grants and are therefore committed to make their research data publicly available after publication of the related article. Accordingly, an institutional repository will be built to provide the researchers with a platform, which can be used for publication of data based on research done at ST Austria, to fulfil the requests from funding agencies (ERC, FWF, HFSP).

## Study

The actual state of the research data at IST Austria was investigated through an on-line survey, distributed via Google Docs. A link to the survey was provided to all professors, who could answer the survey by themselves, or delegate this within their research group.

When the survey was distributed, there were 19 active research groups at IST Austria and all responded. We were looking for one set of answers for every research group. The answers for all questions were defined as free text (no predefined answers), to receive a possibly broad range of answers.

The full text of the survey can be found in the appendix.

# Results

The results are interpreted according to the purpose of the survey - to support the choice of suitable software for publication of research data.

### 1/ How are research data produced in your research process?

Because of the variety of research disciplines, there is a broad range of different procedures: computer simulation, writing code, microscopy, behavioural observation, electrophysiological measurement, DNA sequencing, genomics, in silico simulation, taking pictures, optical density measurement etc.

A full list of the procedures is available in the appendix.

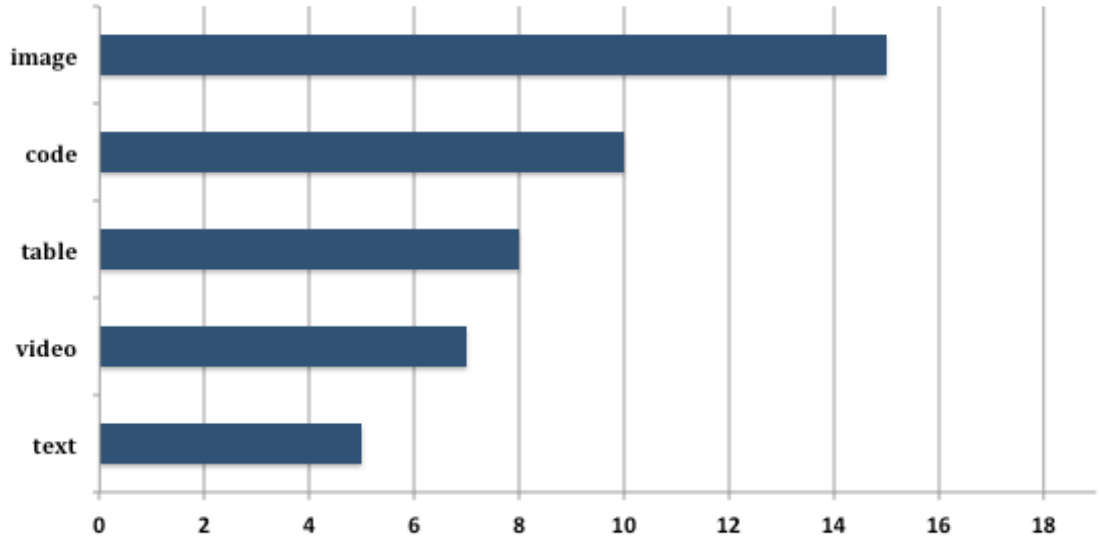
### 2/ Out of which device/system or in which way (for non-digital data) did you get the research data?

Within the 19 research groups, the research data are won through: PC, gas chromatograph, plate reader, patch clamp amplifier, Real Time PCR machine, video camera, photo camera, microscope, GPS, datalogger, CCD camera, fluorescence reader, online genomic database etc.

A full list of the devices and systems is available in the appendix.

### 3/ What kind of research data does your research group produce?

The answers were encoded to 5 main groups of data to simplify the output.
**code** (algorithm, software etc.)
**text** (txt, xml etc.)
**image** (photo, graphic, plot etc.)
**video** (picture stack etc.)
**table** (SPSS, structured textual data, xls)



**4/ What formats of research data does your research group produce?**

Across the research groups, a very broad range of data formats is produced. To see the frequency of occurrence, the formats are ordered according to usage within the research groups.

**data format ( number of research groups, using this format)**
jpg (9)
txt (6)
tiff (6)
eps (3)
pdf (3)
tex (3)
c (3)
cpp (2)
doc (2)

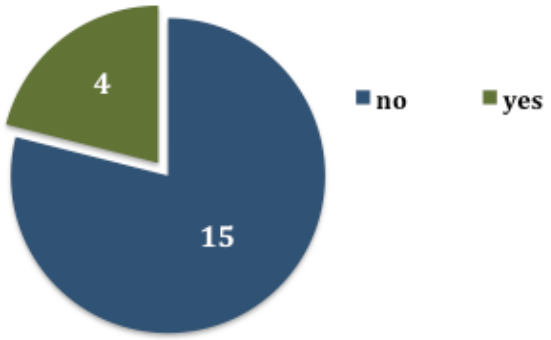
ab1 (1), avi (1), bmp (1), cfs (1), clc (1), csv (1), lif (1), ml (1), mov (1), mp4 (1), mpg (1), nb (1), obj (1), ps (1), py (1), r (1), sav (1), spv (1), stk (2), tikz (1), xls (1), xlsx (1), xml (1)

**Conclusion:**
Formats **jpg**, **txt** and **tiff** are the most used and should therefore be tested first, to ensure proper handling within the repository. All mentioned data formats should be generally supported.

Definitions of listed data formats are available in the appendix.

**5/ Do you have special requirements for describing the research data?**

4 research groups out of 19 would like to add additional description of the data (gene sequences, bacterial strains, brain slices, links to specimens) to the published dataset. 3 of them are from the field of Mathematical and Evolutionary Biology, 1 from Neuroscience.




**Conclusion:**
The repository should enable upload of additional description and materials related to published data.

**6/ What is the approximate amount of research data per month?**

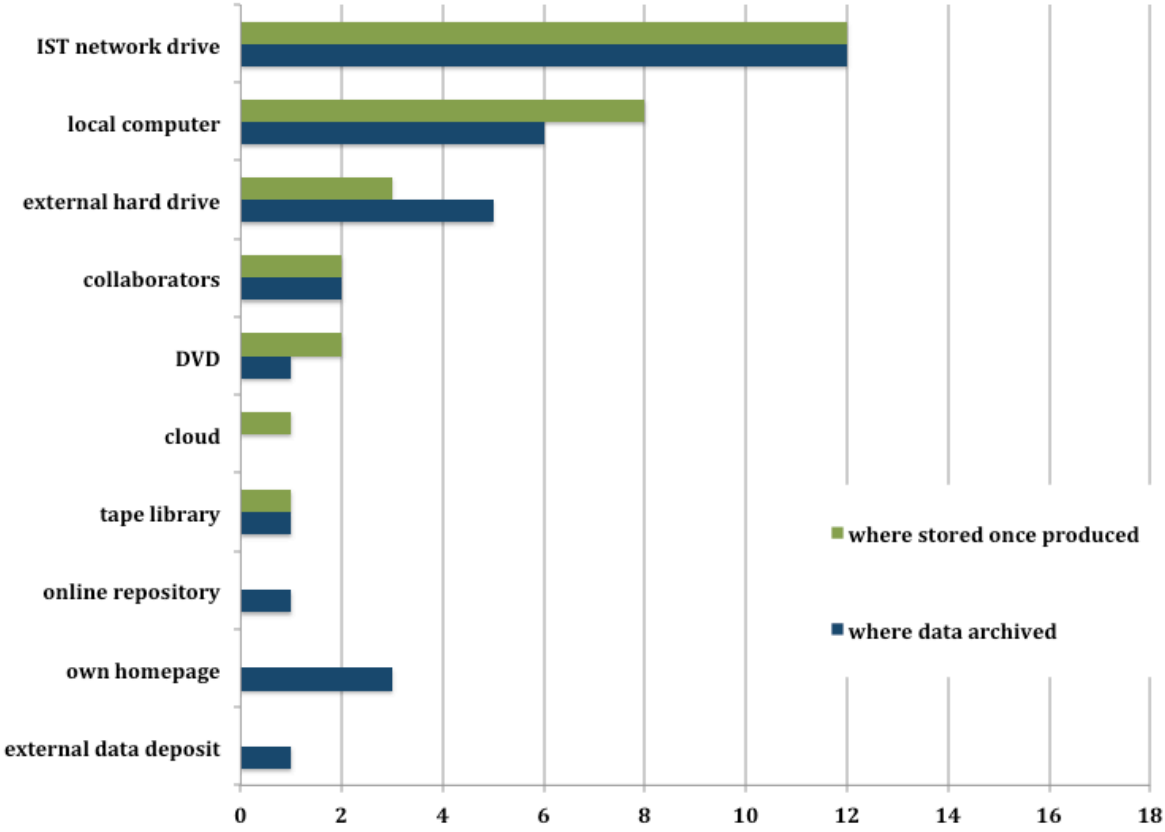
The amount of research data for all 19 research groups together is approximately 1,3 TB in a month.

**7/ What is the approximate average data size?**

Approximate average size varies between a few kb to dozens of GB depending on the research field. Neuroscientists are producing hundreds of GB in a month, Computer scientists only a few MB.

**8/ Where do you store (archive) the research data once they were produced?**
**9/ Where do you store (archive) the research data after publication?**

Representation of results for question 8 and 9 is in the joint table below, which shows the differences in strategies.

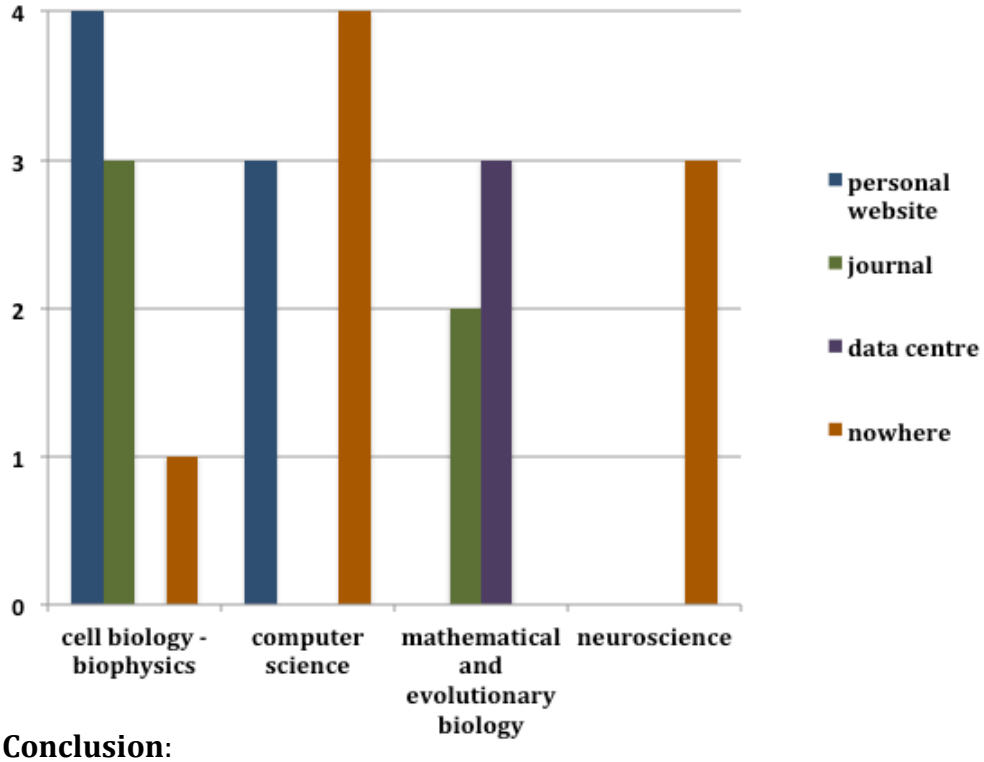



**Conclusion:**
Most of the research groups trust in the IST network drive for temporary data storage and also for data archiving. Still, almost half the research groups use a local computer for data storage, where long-time archiving cannot be ensured. Only a minority of the researches store the data through external methods like data deposits, tape library or cloud.

According to the definition by the DFG[1] in "Safeguarding Good Scientific Practice"[2]: "Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin". This will be better secured by the institutional repository than by individual diverse strategies.

**10/ Where do you upload (publish) your research data after publication?**

The actual experience with publishing data was investigated to see if there is already any experience, or if this is new for the researchers. To simplify the output, the results are grouped by the disciplines to show the difference between the research fields.




**Conclusion**:
Each research field has its own habits for sharing and publishing data, which is visible in the results. In Mathematical and Evolutionary Biology, there is already some experience with publishing research data within journals and also data centres. On the contrary, neuroscientists in our research groups don´t publish their data at all.

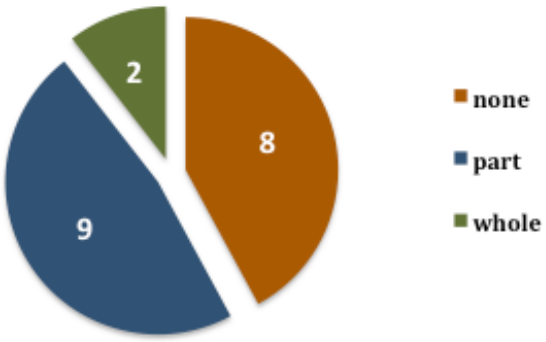
In Mathematical and Evolutionary Biology, the following data centres are used at present for publishing:

| | |
|---|---|
| **GenBank**[3] | DNA sequence data |
| **TreeBASE**[4] | phylogenetic data |
| **Dryad**[5] | other types of data |
| **NCEAS Data Repository**[6] | other types of data from NCEAS funded activities |

[1] Deutsche Forschungsgemeinschaft = German Research Foundation
[2] http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf
[3] http://www.ncbi.nlm.nih.gov/genbank/
[4] http://treebase.org
[5] http://datadryad.org/
[6] http://knb.ecoinformatics.org/knb/style/skins/nceas/

**11/ Do you upload (publish) all of the research data or only a part of them after publication?**

As it is already visible in the answers on the previous question, 8 out of 19 research groups don't publish their research data at the moment. 9 groups publish part of their data (final movie, part of sequence data, supplementary material, rendered video, images) and 2 groups do publish the whole primary data.




## Summary

We started the project „Research Data @ IST Austria" with an on-line survey to investigate the actual state of research data at our institute. Thanks to 100% response rate, we could work with all individual details and needs through all available research disciplines on IST Austria.

As it is visible from the results, it is important to offer researchers the possibility to publish their research data directly in an institutional repository with the opportunity to save links to published data in well-established subject repositories (data centres). At the moment, the individual strategies of the researchers can´t ensure secure storage over a longer time period and also don´t support publication of the data as requested by the funding agencies.

The repository should support the mentioned data formats and also enable upload of additional description and materials related to published data.

In following document, standards, which should be taken into account, will be summarized, as well as existing benchmarking projects and related initiatives and projects.

# Appendix

**Survey text**

**SURVEY to "Open Access to Research Data @ IST Austria" project**

**Please, in this survey understand the term "research data" explicitly as primary data, which are the basis for published documents (e. g. journal article, conference paper, book chapter, dissertation thesis etc.)**

1/ How are research data produced in your research process? (e.g. measuring, microscopy, taking pictures, electrophysiology traces)

2/ Out of which device/system or in which way (for non-digital data) do you get the research data?

3/ What kind of research data does your research group produce? (e.g. pictures, tables, videos, picture stacks, codes)

4/ What formats of research data does your research group produce? (e.g. JPG, TIFF, CFS, LIF format for pictures etc.)

5/ Do you have special requirements for describing the research data? (e.g. biological specimens etc.)

6/ What is the approximate amount of research data in one month?

7/ What is the approximate the average data size?

8/ Where do you store (archive) the research data once they were produced?

9/ Where do you store (archive) the research data after publication?

10/ Where do you upload (publish) your research data after publication?

11/ Do you upload (publish) all[7] of the research data or only a part of them after publication?

[7] All primary data, which are the basis for published documents.

**Full list of procedures**

automated optical density measurement
behaviour observation
biophysical techniques
bulk fluorescence
collecting biological specimens from the wild
collecting observational data of pollinators in the wild
computer simulation
computing results with software
data rendering
DNA sequencing
downloads from public sources on the Internet
electrophysiology
experiments on plants
FACS sorting/counting
gaschromatography-mass spectrometry
generation of simulated data
immunological measures
implementing algorithms
in silico simulations, genomics
measuring
microscopy
photography
processing (feature extraction) using self-written software tools
quantitation
quantitative measurements
reuse of collaborators data
running simulations
single cell microscopy
software tools
source code collecting from open source software
spatial data
taking pictures
writing code

**Full list of devices/systems**

biophysical measurement tool
CCD camera
cluster
datalogger
field
fluorescence reader
gas-chromatograph
GPS
microscope
online genomic databases
patch clamp amplifier
PC
photo camera
plate reader
portable hard drive
public sources from internet
Real Time PCR machine
spinning disc
video camera

**List of data formats**

| | |
|---|---|
| **ab1** | ABIF file format / highly specialized DNA sequencing format |
| **avi** | Audio Video Interleave movie file |
| **bmp** | standard Windows bitmap image |
| **c** | C/C++main source code file format |
| **cfs** | binary format |
| **clc** | UltimaCalc text file |
| **cpp** | C++ main source code file format |
| **csv** | Comma Separated Value file |
| **eps** | Encapsulated PostScript file |
| **jpg** | JPEG bitmap image format file |
| **lif** | Leica image file |
| **ml** | ML language source code file |
| **mov** | Apple QuickTime digital movie file |
| **mp4** | MPEG-4 video file format |
| **mpg** | video file format |
| **nb** | Mathematica notebook file |
| **obj** | 3D object file |
| **pdf** | Adobe Portable document format |
| **ps** | Adobe PostScript file |
| **py** | Python script language source code |
| **r** | source code |
| **sav** | SPSS Data Sets database file |
| **spv** | SPSS (Statistical Package for the Social Sciences) analysis output file |
| **stk** | HyperStudio stacks file |
| **tex** | TeX/LaTeX text document |
| **tex** | TeX/LaTeX text document |
| **tiff** | Tagged Image File Format bitmap image |
| **tikz** | PGF/TikZ is a tandem of languages for producing vector graphics from a geometric/algebraic description |
| **txt** | simple text file |
| **xlsx** | Microsoft Excel 2007/2010 Open XML workbook file |
| **xml** | XML document file |