Version 1.3

Information Services - University of Edinburgh

SCE009 Virtual Edinburgh Recommendations

1 Document Management 4

*1.1* Contributors 4

*1.2* Version Control 4

2 Introduction 6

3 Executive Summary 8

3.1 Virtual Edinburgh Core 8

3.1.1 Stage 1 – Data and API layer 8

3.1.2 Stage 2- Solution repository 9

3.1.3 Stage 3 – Workbenches 9

3.2 Initial Use Cases for delivery 10

4 scale 10

4.1 Recommendations 10

5 Authentication and security 11

5.1 Recommendations 11

6 Open Data 11

6.1 Research Data publishing 12

6.2 Corporate open data 13

6.3 City/External Open Data 13

6.4 Data licensing 13

6.5 Data publishing workflows 14

6.6 Data interoperability 16

6.7 Git 16

6.8 Recommendations 17

7 API Layer 17

7.1 Data Download 18

7.2 API 18

7.3 Loopback 18

7.4 Recommendations 19

8 Solution Repository 19

8.1 Recommendations 20

9 Virtual Workbench 20

9.1 Backup requirements 20

9.2 Recommendations 21

10 References 22

# Document Management

## Contributors

|  |  |  |
| --- | --- | --- |
| **Role** | **Unit** | **Name** |
| Systems Analyst Designer (Owner) | IS Applications | Richard Good |
| Workgroup Leader Software Engineering | EDINA | Ben Butchart |
| Head of Web, Graphics and Interaction | Learning, Teaching and Web | Martin Morrey |
| Project Manager | IS Applications  | Morna Findlay |
| Chair in Technology Enhanced Science Education | School of Biological Sciences | Jonathan Silvertown |

## Version Control

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Version** | **Author** | **Section** | **Amendment** |
| 12/01/16 | 1.0 | RG | **All** | **Initial draft** |
| 13/01/16 | 1.1 | RG | **34.2 7.3 7.6,7.78.19** | **Add component diagramAdd detail to use casesAdd section on external/council dataRemove explicit recommendation of GitHub, add GitLab informationRemove explicit mention of DrupalCombine separate maker/jupyter workbenches into single workbench** |
| 15/01/16 | 1.2 | RG | **3.13.1.13.256.56.66.76.79** | **Update diagram to include geo-location serviceUpdate open data policy recommendationChange intro description to one or more demonstratorsUpdate to indicate potential for demand, and alternative authentication methodsAdd paragraph on selection of GitAdd new workflow for data harvestingAdd new section covering data interoperability.Include Git explanation, and include GitHub pricing link.Update open data policy recommendationAdd link to Maker tool criteria page** |
| 26/01/16 | 1.3 | RG | **All** | **Removed comments.** |

# Introduction

This recommendation report is part of the SCE009 Virtual Edinburgh Scoping Study project. To quote from the project overview:

The aim of Virtual Edinburgh (VE) is to make Edinburgh the Global City of Learning by turning the entire city and its environs into a pervasive, interactive learning environment, visible to the world.

This project will build on the work completed under SCE005, Virtual Edinburgh Business Case Support, by developing the practical aspects of what can be delivered in the following stages, and how this is facilitated across the different parts of IS involved, including EDINA and Service Management.

This project will undertake a scoping exercise to deliver a clear route for how the full Virtual Edinburgh vision can be accomplished, broken into stages that can be accomplished within the funding envelopes available (Stage 2 has Innovation Funding).

This document draws together the SCE005 initial themes into a set of recommendations to take forward into the next foundation stage. The components diagram from the project brief is included for convenience below.



Figure . Virtual Edinburgh components diagram

The next section contains an executive summary of the key recommendations, with following sections providing details on specific components.

A separate project is looking to evaluate and recommend maker tools for creating applications, this document covers mainly the API and data layer in diagram above. For completeness though, the following sections cover:

* API Layer (including upload API)
* Open Data (including schema)
* Authentication/Authorisation
* A solution repository for creators to find out information
* Workbenches providing tools for students and creators to use

# Executive Summary

## Virtual Edinburgh Core



Figure . Virtual Edinburgh component diagram

The core of Virtual Edinburgh provides the necessary components for creators to easily build apps/solutions or make innovative uses of the data available. We recommend splitting the delivery of components into three discrete stages, to be able to start delivering features earlier, and allow the implementation of the initial use cases listed in Section 3.2.

### Stage 1 – Data and API layer

Stage 1 implements the core open data repository and API layer.

The recommendations for this stage are summarised as follows:

* **Cloud first**
* **EASE for web based authentication**
* **SSL encryption for all data transfer**
* **OAuth2 for web based API authorization**
* **Loopback for API production**
* **An official university policy on publishing open data should be created which ties into existing research project planning, and existing corporate data collection**
* **A Virtual Edinburgh group of Git repositories should be created**
* **Data repositories should be created from existing research/corporate data, beginning with data sets which are to be used for the initial use-cases.**
* **Each data repository should be licensed as permissibly as appropriate, e.g. CC-BY, CC-0**
* **Textual data should be stored in a tabular standard format, e.g. CSV**
* **Data in the repository should be clearly documented to ease understanding**
* **People should be able to easily contribute new data, or request amendments to existing open data using standard Git workflows**
* **Data owners, creators should be clearly stated**
* **Data provenance should be clearly stated**
* **Loopback.io will be used to provide APIs**
* **An API should be provided for every open data dataset published**
* **When someone is creating a new app, they will use Loopback for their own internal APIs**
* **For private APIs, OAuth combined with EASE should be used.**

### Stage 2- Solution repository

Stage 2 implements the solution repository site.

The recommendations for this stage are summarised as follows:

* **A solution repository site is set up to provide a one-stop set of guides/solutions and pointers to creating and working with Virtual Edinburgh content**

### Stage 3 – Workbenches

Stage 3 implements the workbenches providing makers and students ready made tools for interacting with data and creating apps.

The recommendations for this stage are summarised as follows:

* **Workbenches running JupyterHub should be made available to students, making use of open data where appropriate**
* **JupyterHub workbooks should be EASE protected**
* **The Maker evaluation project will provide solution for Maker Workbench**

## Initial Use Cases for delivery

For the next stage of the project, we recommend delivering on or more of the following demonstrators:

* Guided Tour App
* Citizen Science survey
* Medical Research Kit app

# scale

The ambition of the project is to create an innovation platform for the whole of Edinburgh. Virtual Edinburgh as a whole has to scale to be able to handle:

* Large datasets
* A large number of datasets
* Large volumes of creators building apps and websites
* Large volumes of people downloading and/or reading the open data in more simple terms, e.g. Excel spreadsheets

## Recommendations

* **It is recommended that the solution is a cloud-first one.**

# Authentication and security

Whilst a great deal of Virtual Edinburgh deals with open data, there are still various aspects which could be considered private, for reasons such as intellectual property, or private research which happens to make use of open data.

EASE is the University of Edinburgh's web login service, providing a Single Sign-On solution for any web-based applications. EASE is also usable with Shibboleth, which allows other trusted organisations (such as other Universities) to sign in using their Single-Sign On solution.

The Authentication method though should be able to be flexible, for example if we later decide to also allow Google+ or Facebook accounts to access maker tools.

To make sure data is transferred securely, we should use an appropriate level of SSL to encrypt data.

Authorisation for each component will be detailed individually.

## Recommendations

* **Any Virtual Edinburgh components which are web facing and require authentication will use EASE**
* **Other partner institutions can use Shibboleth to also provide their own Single Sign-on**
* **All data transfer and sites will use SSL encryption**

**NOTE:** Use of EASE requires either an account to be created in the University Identity Management System, or an EASE Friend account to be created. At this stage it isn’t known how many EASE Friend accounts may end up registering, but there is the potential for this to be high numbers should Virtual Edinburgh prove popular.

# Open Data

Key to Virtual Edinburgh is the ability to consume and publish open data. There are already a number of areas in Edinburgh where open data is available, such as:

* <http://edinburghopendata.info> - The Edinburgh open data portal provided by Edinburgh Council
* <http://datashare.is.ed.ac.uk> - Edinburgh DataShare provided by University of Edinburgh

Although data is already available, it is not necessarily in an easily usable format, nor are there any APIs to allow programmatic consumption.

The [5 Star format for open data](http://5stardata.info/) states the following:

1. **Make the data publicly available**
2. **Make it available as structured data (e.g. excel instead of image scan of a table)**
3. **Use a non-proprietary format (e.g. CSV instead of Excel)**
4. **Use open standards from W3C such as RDF and SPARQL to identify things**
5. **Link your data to other data to provide context**

Realistically most open data repositories aim for 3 stars.

There are a number of very useful resources for detailing how to effectively structure, organize and manage your data, so this report will not detail that aspect.

## Research Data publishing



Figure . Data management planning lifecycle diagram

The University already has a mature planning and organisational process for managing data within research projects. Indeed, research projects are already producing large quantities of data which are available for others to use (although not all data would be considered Open). It should be entirely possible to introduce minor decision points into the existing planning process to facilitate data being made available for Virtual Edinburgh to use, ideally with the data format used to capture data being in a complementary format (e.g. tabular CSV) for immediate use without transform. The idea that a project may produce Open Data of use to Virtual Edinburgh should be considered at the outset of a research project. Meta Data should be present for Research Projects already, so this should be reused and present in the published data.

There will still be effort required though to take the output from a research project and ready it for use within Virtual Edinburgh.

## Corporate open data

The Universities corporate systems store a huge wealth of data, with some of it being able to be published as Open. This can cover a variety of ranges, for example:

* Lists of University courses
* Campus map information
* PC lab availability
* Timetables
* Data on types of food consumed

Data is typically stored in corporate databases, and generally also tied into user-centric information, e.g. a student enrolled on a course.

This data should also be made available as Open Data, although care should be taken to ensure the data is anonymous, and contains no information which can be used to identify people.

## City/External Open Data

As the scope of Virtual Edinburgh is beyond that of the University, it is key to consider the wealth of data which is available in other areas.

Edinburgh Council as previously mentioned has an Open Data portal using CKan with a growing number of data sets added. As of writing, most are in CSV format which is both easier to use and has the added benefit of being able to use the built in CKan data API for querying and using the data. This may avoid us having to build our own API to interact with the read only data. Equally it may allow us opportunity to automatically load our own data repository from the Council to manage changes in the data.

There is a clear opportunity for collaboration and sharing between Virtual Edinburgh and Edinburgh Councils own Open Data efforts, for the benefit of both.

## Data licensing

Data licensing is an important aspect of publishing data, not only to make it clear to someone who is using the data what they can do with it, but also for a publisher to control aspects of attribution for the work they have done in collecting the data in the first place.

For Open Data, typical licensing is fairly permissive, with potential to require that usage includes crediting the creator.

 Licenses mainly fall under creative commons:

* **CC-0** - relinquishes all copyright and similar rights and dedicates those rights to the public domain.
* **CC-BY** - allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited.

## Data publishing workflows

Some typical workflows are created to illustrate the process. In the examples below the publisher/consumer role is separated from the publisher role, as it is possible the two will be separate people.



Figure . New open data creation



Figure . Open data consumed



Figure . Changes to existing open data

The following diagram shows a typical workflow when using data harvesting to monitor for changes in other publishing repositories. Below shows a data owner making a change, a data harvester picking up the change, committing it to the Open Data repository and notifying the Data Publisher. The Data Publisher can then verify the change and publish a new version of the data.



Figure . Data Harvesting diagram

The likely workflows of cloning data, publishing and proposing amendments are very similar to that employed by Git version control workflow.

After reviewing options, Git was proposed to provide the open data repository. Git solutions such as GitHub and GitLab are mature products, with APIs, event driven hooks (i.e. something in the repository has changed), and user friendly visualisations and editors for standard file types which make it easy to interact with data.

## Data interoperability

As some of the usage of data is likely to involve linking data together (data mashups),

Specifications such as Data Catalogue Vocabulary ([DCAT](http://www.w3.org/TR/vocab-dcat/)) are designed specifically to facilitate such interoperability.

In addition, as this is focused more towards APIs and other tools interacting with the data, a machine friendly format such as [JSON-LD](https://www.w3.org/TR/json-ld/) should be used to encode the linked data.

## Git

[Git](https://en.wikipedia.org/wiki/Git_%28software%29) is a widely-used version control system which is primarily used for software development. It is a distributed revision control system with an emphasis on speed, data integrity, and support for distributed, non-linear workflows. It is designed to facilitate changes and history of files in a streamlined efficient manner.

GitHub is a cloud hosted Git platform which is free for open repositories, and already has [a showcase set of repositories used for Open machine readable datasets](https://github.com/showcases/open-data). GitHub file storage is by default limited to 100MB files, and a maximum of 1GB for an entire repository. For large file support Git Large File Storage can be used, although it is no longer free if you go over 1GB per month bandwidth usage. Costing for Git can be found at <https://github.com/pricing>.

GitLab is an alternative when there is the need for onsite hosting. It has many features over and above the standard GitHub set, such as the ability to provide custom branding and landing pages. The Community Edition is free, Enterprise Editions have an associated per user licensing cost with support included.

It is anticipated for Open Data sets that most will be in text format, e.g. CSV, and therefore size considerations will be less of an issue, however support for binary files must also be included.

There is a very useful post outlining the benefits of using Git for open data on the Open Knowledge Foundation blog: <http://blog.okfn.org/2013/07/02/git-and-github-for-data/> .

## Recommendations

* **An official university policy on publishing open data should be created which ties into existing research project planning, and existing corporate data collection**
* **A Virtual Edinburgh group of Git repositories should be created**
* **Data repositories should be created from existing research/corporate data, beginning with data sets which are to be used for the initial use-cases.**
* **Each data repository should be licensed as permissibly as appropriate, e.g. CC-BY, CC-0**
* **Textual data should be stored in a tabular standard format, e.g. CSV**
* **Data in the repository should be clearly documented to ease understanding**
* **People should be able to easily contribute new data, or request amendments to existing open data using standard Git workflows**
* **Data owners, creators should be clearly stated**
* **Data provenance should be clearly stated**

# API Layer

Open Data in and of itself is useful, but it requires easy methods of accessing and being able to use it.

There are two typical methods of interacting with the data:

1. Download and use the data directly (e.g. opening in Excel to view/interact with data)
2. Use an API to access the data programmatically (e.g. when creating an app which uses the data)

## Data Download

For text based data, a tabular format such as CSV should be used. Data should be meaningfully labelled and supporting documentation provided to give people using the data information on what it is.

Having the data available in this format will allow anyone to easily view and interact with the data, and also assist with use-cases whereby the API is not the primary method of interacting with the data.

Data download can be provided easily by downloading the data from GitHub directly.

## API

Choosing the correct API framework is key, as it has to be both easy to access, and also easy for creators of new applications and technologies to create their own. Equally the framework itself has to be flexible enough to cater for the many ways in which it can and will be used.

The following were identified as high level requirements for any API framework to be able to deliver:

* **Must be easy to use**
* **Must be easy to create new APIs**
* **Must be able to scale to high levels of load and usage**
* **Must be able to use a variety of traditional RDBMS e.g. Oracle, SQL Server, MySQL**
* **Must be able to use a flexible authorization model, e.g. use of OAuth2**
* **Must be able to be used by web clients, mobile clients, and machine/desktop clients**

Based on the requirements above, and after a period of evaluation Loopback.io was chosen as the API framework of choice.

## Loopback



Figure . Loopback system diagram

Loopback is an API framework written in Node.js which, by default makes use of a MongoDB database for data storage. It is lightweight, easy to setup and configure, and supports a variety of Authentication and Authorisation models to control access.

In order to provide easy access to the Open Data stored as part of Virtual Edinburgh, there will be an API created per data type. This will allow creators to easily get at the data programmatically when creating new apps and sites.

## Recommendations

* **Loopback.io will be used to provide APIs**
* **An API should be provided for every open data dataset published**
* **When someone is creating a new app, they will use Loopback for their own internal APIs**
* **For private APIs, OAuth combined with EASE should be used.**

# Solution Repository

The solution repository is intended to provide an easy source of how-to guides, sample projects and code snippets for creators to use when creating new projects. People wanting to create apps or work with data should be able to easily find out how they can via guides/tutorials/video content.

A good example of a solution repository is the Open Knowledge Foundation site which is backed by a GitHub set of repositories: <http://data.okfn.org>.

## Recommendations

* **A solution repository site is set up to provide a one-stop set of guides/solutions and pointers to creating and working with Virtual Edinburgh content**

# Virtual Workbench

The Virtual Workbench is where someone wanting to create a new app or site can receive all the tools, APIs and data they need to start work.

The key aims are:

* **To enable quick setup**
* **Allow creative people to begin creating as quickly as possible**
* **To have automatic management to minimise manual setup and updates**
* **To have a secure location for people to create**
* **To allow workbenches to scale to large numbers, e.g. all students may have one**

A workbench containing Jupyter Hub will allow students to have interactive notebooks which allow dynamic running of Python and R code within them. Workbooks also support markdown formatting for display.

For creators, the workbench is concentrated on creating web applications and/or mobile applications. It is anticipated that the separate evaluation project looking at maker tools will recommend a solution, which may or may not require a workbench (e.g. if the tool is cloud based with it’s own workbench). The Criteria used to evaluate the tools is currently held at the following location: [https://www.wiki.ed.ac.uk/display/VE/Application+Development+Frameworks](https://www.wiki.ed.ac.uk/display/VE/Application%2BDevelopment%2BFrameworks).

Depending on the tool selected, there may still be a requirement to provide API and data hosting in a workbench for creators to use. The maker project should provide the recommendation in this area.

## Backup requirements

Where people are creating and storing information in workspaces, appropriate backups should be in place to avoid them losing data should a disaster occur (such as loss of disk).

Normally, for research projects it is the responsibility of the person collecting the data to ensure that they employ appropriate steps to ensure backups are taken to avoid data loss.

## Recommendations

* **Workbenches running JupyterHub should be made available to students, making use of open data where appropriate**
* **JupyterHub workbooks should be EASE protected**
* **The Maker evaluation project will provide solution and suggest integrations with Virtual Workbench**

# References

1. <http://loopback.io/resources/#compare> - Loopback framework comparison
2. <http://jupyter.org> - Project Jupiter
3. <http://www.ed.ac.uk/information-services/research-support/data-management> - Research Data Management at the University of Edinburgh
4. <http://data-archive.ac.uk> - The UK data archive
5. <http://datalib.edina.ac.uk> - Data Library at Edinburgh University
6. <https://www.era.lib.ed.ac.uk> - Edinburgh Research Archive
7. <http://www.w3.org/DesignIssues/LinkedData.html> - Linked Data, Tim Berners Lee
8. <http://www.ed.ac.uk/information-services/computing/computing-infrastructure/authentication-authorisation/ease> - EASE documentation
9. <http://shibboleth.net> - Shibboleth documentation
10. <http://5stardata.info> - 5 star scheme for open data
11. <https://github.com/showcases/open-data> - GitHub open data showcases
12. [https://en.wikipedia.org/wiki/Git\_(software)](https://en.wikipedia.org/wiki/Git_%28software%29) - Git Wikipedia entry
13. <https://www.w3.org/TR/json-ld/> - JSON LD specification
14. <http://www.w3.org/TR/vocab-dcat/> - DCAT specification