**KIP-INCA Workshop on developing a shared spatial data platform**

**EEA, 9 March 2018**

**Background paper, draft 1**

**1. Introduction**

The KIP INCA 2016 roadmap states in its chapter 6 on ‘Data infrastructure for EU ecosystem accounts’:

‘Ecosystem accounting depends on the availability of geospatial data that accurately describe the distribution and condition of ecosystems with sufficient resolution to capture small ecosystems. This means that a further investment is required at EU level to develop a shared data platform for the integration of ecosystem-related data at large spatial scale (<1:100 000) than currently feasible for a number of data sets. Such a shared data platform would be a key building block for operational EU ecosystem accounts but would also provide substantial value-added for other analytical purposes.’

In the context of this objective, the KIP INCA partners have to agree on what elements of such a data platform should be developed and how far these elements should be shared. A key element is a shared structured approach that enables harmonized, quality controlled and quality assured storage, access and sharing of spatial data while allowing targeted search by environmental topics. Further elements of a spatial data architecture could address, among others, automation of work flows related to geo-processing scripts and models, servers, publishing and data integration, editing environments, and semantic inventories.

**Key tenets for developing a shared spatial data platform under KIP INCA (taken from 2017 paper):**

* Aim is to develop a shared platform that is stable, ‘institutional’ and achieves full integration of systems and sharing data sets (as far as appropriate)
* Develop accounts underpinned by harmonised, geo-referenced and QA/QCd datasets
* Develop a semantic inventory of spatial datasets placing data in their context.
* Serve as common spatial reference framework; with a 1km grid as the core accounting grid
* Ensure transparency, traceability and accountability of the datasets and models used to compute the accounts
* Facilitate greater efficiency in sharing of spatial data and related joint analysis, thus enabling a pooling of resources between KIP INCA partners and synergies
* Achieve this making best use of the present capacity and resources

**Key issues and questions to consider during the workshop:**

There are two main areas to consider when reflecting on developing a shared spatial data platform:

a) Efficient sharing of existing data sets in the context of ‘regular production’

b) Development of shared spatial reference data on natural capital

For a) we need to know:

* Do we have a problem? If so, how can it be fixed? Within existing tools and set-up or do we need to create a new technical solution?
* Also need to consider context: experimentation or regular production?

For b) we need to agree:

* Which are the priority reference data sets on natural capital
* Can we create them by joining forces? How to set that up and make it a regular activity?
* Are new monitoring or data collection schemes necessary?

**Reflection on data foundation for experimentation and for regular production** – there are some key differences – see table below:

|  |  |  |
| --- | --- | --- |
| **Issue** | **Experimentation** | **Regular production** |
| **Main objective** | **Demonstrate feasibility**  | **Efficient & stable production according to agreed methodology** |
| **Data source** | **Grab and use what is available** | **Go for ‘official’ and regular data** |
| **Data integration / alignment (between ecosystem accounts)** | **Not so important as long as analytical tools can cope** | **Full alignment supports efficient production, enhances credibility of ecosystem accounting system** |
| **Production approach** | **Tailor to analytical purpose, can be flexible, can be one-off** | **Requires automated production as much as possible, needs to establish a regular and fully comparable time series** |
| **Institutional set-up** | **As necessary, voluntary arrangement suffices** | **Formal cooperation required; full compatibility of systems and standards** |

**Key questions to review in relation to data platform functionalities**

1. Who shares data with whom? Is current set-up efficient for data access?
2. If there are obstacles what are these? How to solve them?
3. How do we ensure common geometry and other key data standards for efficient data sharing and integration into different analytical tools.
4. How to ensure there is one master version which everyone uses, how is that maintained/updated?
5. Are there output data sets created by JRC/EEA others that are not officially published? How to deal with / efficiently share these?

We should look into the set-up of EEA IDP project once more to identify key issues for creating and maintaining a common pool of data to be used by various partners. However, we need to be aware that the tools combined under the EEA IDP project are currently focused on use within the ‘EEA family’ – hence opening them up to INCA partners may be difficult; or at least require substantial investment!

Some final points to take account of:

* Consider link to review of EU Env. data centres (EDCs) – KIP INCA is a ‘use case’ and some of the INCA needs may be resolved via agreements achieved in the EDC review process
* An ‘EEA IDP like’ system may not be absolutely required for access to, or sharing of, existing official data that all partners push out via EDCs
* However, establishing a shared data platform based on the tools and principles developed in the EEA IDP project may still yield substantial efficiency gains and ensure that we use the same version of input data sets used across different INCA components
* We need to establish a system/approach that ensures that NCA reference data created via or for KIP INCA are properly maintained, version controlled, and accessible via a ‘one-stop’ system
* ‘Governance ‘ ?
* Resources ?

**2. State of play at key KIP INCA partners**

EEA:

EEA’s spatial data infrastructure (SDI) is already in place, INSPIRE compliant and stores large amount of European datasets and has been tested for spatial analysis of the EEA and its European Topic Centres. Given that it is a very suitable platform for supporting all KIP INCA partners in work relating to spatial data integration and spatial analysis. The catalogue interface of EEA’s SDI (<http://sdi.eea.europa.eu/> ) facilitates fast and effective data sharing. Other elements of a spatial data architecture are also in place in EEA’s system architecture while additional tools for harvesting and processing data sets have also been developed at the JRC. It is however expected that certain data analyses are best performed using KIP INCA partners own systems reflecting best practices developed for their specific analytical environments. Therefore, many spatial data analyses are likely to be performed at the KIP INCA partners’ local premises, using already available tools while staying connected to, and benefitting from, a common spatial data infrastructure.

EEA runs a project called Integrated Data Platform (IDP) which can guide the development of the INCA shared data architecture. The project supports the integrated analysis and assessment of geospatial data. It ensures populating EEA’s SDI with harmonized, quality controlled and quality assured spatial data, and offers advanced options (semantic inventory and network analysis) for querying and understanding the datasets. The metadata of these datasets can be queried in the SDI metadata catalogue. See more info [here](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353808/EEA%20IntegratedDataPlatform.pdf?api=v2). Additionally, the beta version of EEA’s so-called JEDI (Joint Environmental Data Infrastructure) toolset, a web-based system, is now available. This performs spatial data integration and conversion into a tabular format ready for computing statistics (using the Tableau software environment).

JRC:

JRC has a metadata catalogue (<http://data.jrc.ec.europa.eu/> ) which describes the available data sets at JRC and provides a link to where access to the data can be requested. See more info [here](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353808/JRC-DataPolicy-INCA.PDF?api=v2). The catalogue is considered compatible with the EEA’s as both are compliant with European standards (INSPIRE). However, further review and likely investment in IT developments are still necessary in order to ensure that the two catalogues can fully communicate. It is to be noted that while JRC’s catalogue stores only the information where to find spatial datasets, the EEA’s catalogue is built upon and references a Spatial Data Infrastructure which physically stores harmonized and INSPIRES compliant spatial datasets. On one hand this ensures that all EEA’s systems communicate and hence the harvesting of semantic information of the spatial datasets can be automatized and inventoried including updates and versioning. On the other hand the SDI ensures a harmonized spatial data pool with only a minimum need for further data processing (e.g. scale adjustments). This makes the cooperation of different parties in shared spatial analysis (so far the EEA and its European Topic Centres) much more efficient and reliable.

The JRC tools DOPA and EASIN are two online platforms which connect to global and European data providers (outside the Commission), process these data and publish this processed information for specific purposes (see links to the [DOPA](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353808/DOPA%20architecture.pdf?api=v2) and [EASIN](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353808/EASIN%20architecture.pdf?api=v2) presentations). The difference to EEA’s JEDI system is that JEDI ensures the automatic integration of geospatial data and conversion of these data into a tabular environment.

[*2017 reflection, possibly outdated now*:] On the other hand, DOPA and EASIN offer spatial data processing and presentation capabilities. Hence the development of the KIP INCA data architecture could consider which role DOPA/EASIN can play in combination with EEA’s SDI and JEDI system. This may focus on capturing data that do not correspond to EU INSPIRE standards (as these tools have been designed to deal with a wide variety of data) or on the presentation of data and analytical results.

Eurostat:

To be completed..

**3. Lessons learned for KIP INCA from discussions in late 2016 and early 2017**

Eurostat summary of meeting on 1 February 2017:

**Minutes provided by Eurostat on KIP INCA data architecture discussion:**

* Solid [proposal](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/189268362/INCA%20Data%20Architecture%20KIP%20Proposal.pdf?version=1&modificationDate=1486629202742&api=v2) for architecture,largely based on existing EEA data infrastructure (EEA SDI and EEA integrated data platform);
* JRC will continue to manage data and models for their purposes but the central data management capacity will be with the EEA, supported by ETC;
* No accounting specific data and metadata standards seem to exist, as a result, the proposed data and metadata standards are EEA extensions of INSPIRE, and are harmonised by EEA/topic centres before integration into INCA by EEA, this makes the architecture dependant of ETC resources;
* Strict guidelines on semantics, documentation practices, reference versions and versioning will need to be developed (case for  linked data);
* Interoperability of JRC and Eurostat data and geospatial data catalogues needs to be ensured by adhering to standards (mainly INSPIRE for data and metadata and DCAT for metadata in the JRC);
* Data architecture is mainly a process description and description of where to find data, a common and formal process description standard needs to be adopted, e.g. based on ETC practices;
* The architecture should be presented to Member States to have their buy in as they would have to use similar  approaches for their national data to ensure interoperability; We need to find an occasion for this presentation.

**Decision**

**The EEA proposal (architecture building on EEA SDI, EEA integrated data platform - data processes building on established ETC work flows) represents a good compromise between distributed architecture and control on data structure, semantics and metadata). The proposal will therefore be the candidate for building the INCA data data infrastructure. Demonstrating the feasibility will be one of the priorities for 2017 and hence one of the main goals for the showcase task in 2017, to be completed by Quarter 3 2017.**

Points summarised in background paper for that 2017 meeting (partly outdated now):

* The development of KIP INCA data architecture should build on the current tools, technologies, and resources to the maximum extent possible. However, for efficiency the different systems need to communicate, which implies that some non-substantial but necessary IT development may be needed.
* Processing of data should in most cases be done locally, at each partner’s premises using best practice processing environments but should use harmonised and standardised spatial data sets. This means that efforts should mainly go to harmonisation and standardisation of data models and their meta-data so that KIP INCA partners working on calculating natural capital accounts can find and download the datasets on a local drive. The results of data analyses, development of accounts, maps and indicators should be made available on a joint landing page for the KIP INCA project.
* The development of a shared catalogue is recommended for distributed systems such as INCA. The experience of EASIN teaches us that developing a shared catalogue which describes meta-data drives standardisation and harmonisation. The compatibility of current catalogues at EEA, Eurostat and JRC for this purpose needs to be further reviewed. To begin work on this task it is proposed that EEA’s SDI is populated with key KIP INCA reference data sets, using available ETC ULS resources in 2017.
* Spatial data harmonization is still extra further step which requires extra resources, as explained under the previous point. In order to support this task EEA’s IDP project has developed a QA/QC protocol listing criteria and setting up processes for spatial data delivery.
* EEA’s Integrated Data Platform develops a semantic inventory of spatial data which enables the fast identification of those datasets that have the appropriate technical and contextual detail for analyses. A subset of this inventory system could be developed for spatial datasets used in the various KIP INCA modules.
* EASIN allows users to trace back the data that have been used to obtain a certain outcome.

**4. 2017 proposal for the KIP-INCA shared data architecture version 0 (KIP-ISDA-v0)**

The following proposal describes a possible workflow for developing natural capital accounts (extent, condition, services). This reflects the state of art in February 2017 and some ideas are no longer being pursued, e.g. the use of the DOPA explorer for the presentation of results on a KIP INCA landing page.

**The core of KIP-ISDA-v0 are two catalogues**: one for geo-referenced, harmonised datasets and one for models. We suggest to develop such a catalogue based on experiences of JRC and EEA but with a joint KIP-INCA visual identity. The KIP-INCA metadata and meta model catalogues can be simple copies of the relevant metadata present in the EEA SDI and JRC meta model catalogues. The two INCA catalogues contain the descriptions of data and models which are necessary to calculate accounts whereas EEA’s SDI physically stores the harmonized spatial datasets. EEA’s semantic inventory harvests its SDI catalogues and enables the fast identification of those datasets which may be used in the same accounting calculations. Accounts are calculated by an operator on a local drive accessing a spatial data infrastructure (for example EEA’s SDI) through the semantic inventory (such as that of the EEA IDP), and final results are uploaded on an INCA landing page. Some IT development is expected necessary for the presentation of results on a KIP INCA landing page, a possible candidate tool for that purpose being JRC’s DOPA explorer.

The figure below shows the diagram developed in early 2017 for reflecting on developing a shared ‘KIP INCA data architecture’.



*Note: The figure above shows February 2017 proposal for the KIP INCA data architecture. The beige shading indicates the components that we consider the ‘data architecture’ proper. The components outside it can be considered part of a wider system architecture. EEA outputs under MAES ecosystem assessment and accounting related products are already being integrated in the EEA-SDI.*

*Several issues still need to be discussed, e.g. how the metadata catalogues for JRC and Eurostat fit into the KIP INCA data architecture.*

The following work streams are recognised (see the numbers in circles):

(0) - (1): ***Geo-referencing and harmonisation of spatial data sets:*** In order to integrate geospatial data for accounting spatial harmonization work is necessary. This is already ensured in the spatial dataset registered in EEA’s SDI with a common geographic reference system as well as a harmonized and controlled spatial geometry and spatial extent. If the goal is a shared KIP INCA data pool where all partners can use all datasets then ideally JRC’s, ESTAT’s and EEA’s datasets for KIP INCA should be harmonized and stored in the same spatial data infrastructure, optimally in EEA’s SDI. Only if all KIP INCA datasets are stored in the same spatial data infrastructure can harmonization, maintenance and update of the datasets be ensured. The metadata referring to the spatial datasets which are relevant for INCA are flagged in the semantic inventory module of the IDP, which harvests EEA’s SDI. Overall the Integrated Data Platform coordinates the smooth population of the catalogue and of the semantic inventory. The harmonization, quality assurance, SDI registration and inventory of spatial datasets will not require IT development but will require additional human resources.

(2): ***Data sharing or integration:*** JRC data sets for INCA are stored on a JRC data server and can be consulted via the JRC data catalogue. The metadata referring to the datasets which are relevant for INCA are flagged and these metadata are collected on the shared INCA metadata catalogue. *Please note that this description refers to a situation where data are retained in local systems and thus do not become freely or easily accessible to all partners.*

(3): ***Data sharing or integration:*** ESTAT data sets for INCA are stored on an ESTAT data server and can be consulted via the ESTAT data catalogue. The metadata referring to the datasets which are relevant for INCA are flagged and these metadata are collected on the shared INCA metadata catalogue. *Please note that this description refers to a situation where data are retained in local systems and thus do not become freely or easily accessible to all partners.*

(4): ***Harvesting and integration of data sets from non-EU and/or not standardised sources:*** INCA will rely on other datasets as well (e.g., species distribution data from IUCN, data from UNEP or FAO, alternative land cover land use data sets; earth observation data managed by global or European data providers, ...). DOPA’s technology could be used for processing the datasets but their harmonization, quality control, maintenance and updates are not ensured. Therefore for those spatial datasets which are expected to be used in processes on a regular basis the SDI registration, and thereby harmonization ensuring comparability, may be the optimal way forward. EEA has approx. 100 man days in 2017 for developing the KIP-INCA data foundation including identification, collection, harmonization, SDI registration and semantic inventory of those spatial datasets which will be used on a regular basis.

(5): ***Selection and documentation of models employed:*** Most ecosystem services accounts will probably require some degree of modelling. Direct measurements of ecosystem service flows are rare, certainly on regional to global scales. The JRC operates several models which are useful for INCA (see Annex 4 of the INCA report). These models are in principle documented in [MIDAS](http://midas-portal.jrc.it/discovery/midas/). However, it seems that MIDAS is only accessible from inside the European Commission. This would mean that we have to manually copy model documentation (e.g., via a pdf print) to the INCA meta model library. MIDAS is a model documentation system but the JRC has no model library where users can download all models for personal use. JRC will have to create such a model library for INCA but at this stage it is absolutely not certain if all models can be published under an open access or open data license. The first version of the model library could contain the pollination and recreation models that will be used for the 2017 working program as well as other model codes that are used (for instance algorithms to downscale national statistics).

(6): ***Sourcing of additional modelling tools:*** Many other models, methodologies and tools are available to quantify flows of ecosystem services. Invest for instance is a well-known model set developed by Stanford University (natural capital project). ESMERALDA is currently listing different models, methods and tools and sorts them per ecosystem type, ecosystem services, geographical scale, and tier (1 to 3).

(7): ***Data processing and analysis:*** Data and models are subsequently brought together to quantify extent and condition and physical and monetary flow of ecosystem services. This work needs to be done on local servers, based on the harmonised data sets in the EEA SDI ‘INCA module’, using the partners’ specialised expertise and following guidance and manuals.

(8): ***Publication of maps, indicators and accounts:*** The output of the processing step are maps, indicators, and accounting tables. The spatial data can be stored and documented on the EEA’s SDI following work streams 1, or potentially on JRC and Eurostat servers if working steps 2 and 3 are designed as described above, while links to the data and metadata can be provided on the INCA metadata and metamodel catalogues and indirectly via the INCA landing page. Indicators and accounts can be published on the INCA landing page, e.g. on BISE.

**Annex 1:**

Documents /analysis developed in previous years:

The KIP INCA partners agreed to draft a first proposal for the INCA data architecture by February 2017 with EEA being in the lead for formulating specifications and requirements. The following elements are available

* Background paper of 25 January 2017
* [Reviewing the data foundation for ‘accounting for natural capital and ecosystem services](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353672/Data%20foundation%20for%20natural%20capital%20accounting_INCA%20KIP_resource%20document_Oct%202015.pdf?api=v2)’
* A first proposal for streamlining the INCA data architecture is included in the INCA design report (chapter 6)
* [Short description](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353672/Integrated%20Data%20Platform%20%2B%20KIP%20INCA_V%200.1_14-11-16.pdf?api=v2) of the EEA’s Integrated Data Platform
* Support to the EEA Technical Report on methodological approaches for land degradation: [Extract of annexes](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353672/Land%20degradation%20exploratory%20study_annexes%20on%20data%20sets.pdf?api=v2) on data sets used
* A [list](https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/175353672/Data%20sets%20on%20%27natural%20capital%27_review_LisaW%20%26%20JEP_rev%2008-09-15.ods?api=v2) of datasets on natural capital

**Annex 2: Updated document on ‘data architecture test case’ on spatial nutrient accounts**

[ *to be completed; template for review of data sources and other input for test cases provided as separate document* ]