SERA
Seismology and Earthquake Engineering
Research Infrastructure Alliance for Europe

DELIVERABLE

D5.2 Position Document on Future DSS Data Accessibility

Work package	WP5
Lead	UU, University of Uppsala
Authors	Roland Roberts, UU Ramon Carbonell, CSIC Angeliki Adamaki, UU Monika Ivandic, UU
Reviewers	Management Board
Approval	Management Board
Status	Draft
Dissemination level	Public
Delivery deadline	
Submission date	16.04.2018
Intranet path	DOCUMENTS/DELIVERABLES/SERA5.2_ Position_Doc_on_Future_DSS_Data_Accessibility

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730900.

Table of Contents

The purpose of this document

Within the SERA project, one work package is to investigate an appropriate model for integrating Deep Seismic Sounding (DSS) data into the EPOS framework. EPOS is the "European Plate Observing System", and is a project sponsored by ESFRI, the "European Strategy Forum for Research Infrastructure". This document explains the current status of the work within SERA and presents the current vision of how an EPOS function for DSS data might best be structured. The SERA working group has communicated with the DSS community in various ways, most recently by holding a discussion meeting at the EGU meeting in Vienna (April 2018). Continued dialogue with the community will be necessary to further define the most appropriate way forward. This document is intended to define what has been identified as probably the most sensible structure for DSS management, and to provide a basis for more detailed technical discussions.

This interim report has been produced by workers within the working group, primarily from the University of Uppsala and CSIC in Barcelona. Contributors include Roland Roberts, Ramon Carbonell, Angeliki Adamaki, Monika Ivandic and others.

1 DSS Data

Many seismic methods penetrate the "deep" Earth. In our context here by "DSS" we mean controlled-source seismic data, usually collected along profiles. The term "DSS" is sometimes used to mean only studies with large offsets between shots and the more distance receivers used. Such data is sometimes referred to as "wide angle" data. Such data contains phases reflected from boundaries within the earth, and "diving" waves which are refracted such that they follow curved paths to considerable depths within the Earth and back to the surface. Rays for diving waves and wide-angle reflections propagate over considerable horizontal distances. Depth of penetration is limited by the size of the source and the geographical extent of the receiver arrays. Generally speaking, penetration depth may be from a few kilometres (upper crust) to mantle depths. Array lengths may be tens or hundreds of kilometers, and in some cases even longer. Many such DSS projects have been carried out on a completely non-commercial, research basis.

Collecting such DSS data, especially for longer profiles, is a major task. It would be difficult or impossible to repeat some earlier projects, partly because of difficulties in permission for using the large explosive sources necessary. Therefore, even data of considerable age may be very relevant for future research, and the intention is to design an e-infrastructure which is well-adapted to dealing with both recent and older data, which may exist in different forms.

DSS data contains even near-vertical incidence reflections. Many studies collecting only near-surface incidence ("reflection") data penetrated to the relevant depth ranges. Very large amounts of reflection data exist. Much of this data has been collected commercially, implying significant confidentiality issues, although in the longer term most of the data is likely to be made more freely available. In contrast to wide-angle data, the ray paths involved for reflection data are generally close to the vertical. Because of the very large volumes of reflection data which exist, and that much of this is commercially collected and may therefore not be freely available for research, the SERA working group was hesitant about fully including reflection data in the current work. Instead of considering initiating development of a system capable of dealing with all reflection data, it was considered more appropriate to focus on the wide

angle data, but with the intention that the design of the system for DSS data to be well-suited to a system for reflection data, when this is later fully developed.

However, feedback from the community is that they would like the system to include reflection data from an early stage.

2 EPOS

A “classic” DSS experimental configuration consists of a long profile of recording stations (seismometers), extending over tens or hundreds of kilometres, or more. Station spacings vary from project to project, being steered by the length of the profile and the number of available sensors and their associated equipment for recording the signals. Separations of some kilometers are not unusual for longer profiles. At several points along the profile, shot points are defined, and large explosions are fired at these locations one or several times. Repeating shots at the same location allows for redeployment of recording stations, allowing more recording points along the profile for the given shot point(s). The explosions may be in boreholes, but for cost and logistical reasons are often in water (lakes or the sea). The explosions used may be very large – up to several tonnes of explosives. Even nuclear explosions have been used as DSS sources. It is also possible to use airgun sources. This has the advantage that many close-lying source locations can be use i.e. that there is a source “array” as well as a receiver array. Disadvantages include cost and that the sources can only be where there is accessible water of an appropriate depth for the boat. If boat-driven airguns are used as sources in a DSS study, it may be appropriate also to use a steamer containing sensors (hydrophones) to also record near-vertical incidence data. Also signals from other sources, such as vibrators, can effectively penetrate to considerable depths, if the vibrations are of sufficient amplitude and the Earth structure in the area is appropriate. The amplitude of signals which is practically feasible to generate from vibrators and airguns is limited, and large explosions may be necessary to produce observable signals at greater distances, corresponding to waves sampling the Earth at greater depths (deep into the mantle). In many areas, for e.g. environmental reasons it has become increasingly difficult to get permission to detonate such large explosions, precluding many conceivable new classic long range refraction studies. This means that much of the older data may be regarded as unique and not practically reproducible, so securing data from these projects may be important for the future, despite the major instrumental improvements which have been achieved over the last few decades.

2.1 DSS metadata in EPOS

For DSS data to be made readily accessible via EPOS, EPOS must have the necessary metadata information. This will include information on profile and source locations, locations of recording stations, names of data-collection projects etc. For DSS data, it will also be appropriate to offer access to derived data, from filtered reduced time sections to derived Earth velocity models. Consistent metadata structures for all relevant types of data should be defined. These structures should be fully in-line with the metadata philosophy and structures within EPOS as a whole.

2.2 Access to raw and derived data via EPOS: Thematic core services?

DSS raw and derived data today exists in many different forms at a large number of different institutions around Europe. Because of the nature of DSS projects, total volumes of data are not large compared to some other forms of data, but are nevertheless considerable. At least at this stage, it does not appear to be appropriate to envisage a model where all DSS data is stored by the EPOS ICS. It may, however, be appropriate to consider if some specific types of data should perhaps be stored by EPOS itself. This could include e.g. some derived data such as images of Earth velocity cross sections, or some data which might otherwise be lost for the future e.g. because the institution owning the data will no longer exist.

When designing a possible model for DSS data access via EPOS, it is important to consider if a TCS coordination function specifically for DSS data will be necessary, and if so, what functionalities this should include. Today, a number of European institutions run well-structured databases including DSS data. In principle, one of these, or a new common database, could collect and administer all DSS data. However, this does not appear to be the most attractive model. Instead, it would appear more appropriate that these existing well-structured databases continue to operate, and collaborate with EPOS such that the data stored is readily accessible via the EPOS ICS system. For data in such databases to be discoverable, metadata must be provided to the EPOS ICS. There are different possibilities for how this might be achieved. One is that each database actively imports into the EPOS system metadata relating to (newly available) data. Another is that there is "discoverable" metadata available in a defined manner at each database, and an automatic or semi-automatic ICS function imports this metadata to the ICS system. At least initially, it seems likely that the most appropriate model is that the data owners should actively supply metadata to the ICS. This will require well-functioning software within the ICS, making this supply of metadata as straightforward as possible. A TCS functionality to do, or facilitate, this import of metadata is conceivable. However, it is not clear if this would be a particularly effective way of achieving metadata import, given that it is the individual data owners who have the information regarding their own data sets.

Some DSS raw data exists on storage media (tapes, CDs…) or computer files, but not as part of a structured database. There is often a risk that such data be lost for the future, either through the data itself being lost, the associated metadata being lost, or full information relating the metadata to the data files being lost. It is desirable to secure for the future as much of this data as reasonably possible. One possibility is that each data-owning institution establishes a long-term maintained database structure consistent with, and accessible to, EPOS. In many cases, this should be possible and appropriate. In other cases, data owners may not have the capacity to do this. In some such cases it may be appropriate for the data owning institution to choose to request that some other institution with a functioning database system accepts the data and agrees to maintain this long-term. It would appear that any such arrangements must build upon bilateral agreement between the two parties involved. In some cases, it may not be possible to make e.g. existing raw data directly accessible via the EPOS system. In such cases, including some metadata information in the EPOS database is still desirable, even if possible access to actual data would then demand contacting the data-owning institution and requesting data access.

In discussion with the DSS community, it became clear that there are a considerable number of institutions who have data stored on older media, such as 9 track tapes, but may lack facilities to read this data, if it is recoverable at all. Therefore, the SERA working group will now attempt to coordinate information about where suitable equipment and expertise for reading older media may exist, to assist the community in saving the data which can be saved.

Especially for older data, which may have been collected decades ago, raw data may no longer exist. In some cases, data never existed in digital form. In these cases, only derived data, such as images of reduced time sections, or derived earth models, may be available. Some such derived data exists in digital form at various institutions, some may only exist in the form of e.g. images in published works,

and in others the data may only exist in the form of paper plots at some institution. Providing metadata information to EPOS on such data is desirable. In some cases, it will be natural and straightforward for data-owning institutions to supply metadata information. In others, this may be more difficult e.g. because the data-owning institution no longer exists. One component in dealing with this issue is that there should be a "project database" within the EPOS metadata system. In most cases, this metadata will allow direct access to raw and derived data. In other cases, the only existing data may in the form of e.g. printed sections in published articles. It would appear appropriate for the DSS community to attempt to ensure that such a project database is as complete as can sensibly be achieved. As a first step in this direction, the SERA project is collating a list of DSS projects, as far as possible with information regarding data-owning institutions, published data and results etc (Appendix 1). It is not envisaged that the SERA project will succeed in making this list complete, but it is hoped that the list developed can act as a starting point, to be complemented as time progresses. The information provided in Appendix 1 can and will be complemented during the coming months, as the DSS community supplies SERA with more information. The projects are listed according to the leading institute or relevant data centres. The information such as data location and availability, contacts, published data and results, and, where possible, data format, storage media, etc. is also provided.

2.3 Coordination and administration

One possibility is that some centralized TCS function be established for management of DSS data, delivery of metadata to EPOS, and facilitation of access to data. Long-term financing of such functions may be problematical. Thus, while this is a possibility, some less centralized form of coordination may be more appropriate for DSS data.

For the EPOS DSS system to function, clearly defined and appropriate metadata definitions must exist. These should be adequate for both raw and derived data. Appropriate metadata structures exist at various databases. These are not necessarily identical in structure, may not be well-adapted to the centrally-defined EPOS metadata structures, and may not include metadata structures for all relevant types of raw and derived data. Therefore, metadata definitions and structures will ultimately need to be defined in collaboration with the DSS community and with EPOS, in order to ensure full compatibility with the EPOS metadata philosophy and structures. However, at this stage, the SERA project has listed examples/suggestions of metadata that are already used in several existing DSS databases, such as OpenFIRE (Finland), CSIC (Spain), and BIRPS (UK) (Appendix 2).

It seems likely that a small but representative group of participants from the DSS community should participate in a discussion leading towards definition of metadata structures. Even in the longer term, it seems likely that updates of metadata definitions and structures may be necessary, and thus some long-term function to deal with this would appear to be necessary. This can be regarded as a "TCS" function, but is of limited scope and thus cost. This function could conceivable take the form e.g. of a small "expert panel" directly attached to the EPOS ICS, as opposed to an external organ collaborating with EPOS in TCS-form.

For some envisaged EPOS TCSs, data quality assurance, quality stamping etc is regarded as an important function. Given the nature of DSS data, this is perhaps less of an issue for the DSS community than for some other communities, and it is not self-apparent that a TCS-centralised data quality functionality is necessary or desirable. It will, however, be important that metadata definitions allow for full information regarding possible deficiencies in data (e.g. station location errors, especially for older data, known/possible problems with timing at some recording stations, etc).

Some DSS projects have included data collection by a number of different institutions. In most cases, data has been collected at a single location, and can presumably be made accessible via the host institution. However, issues of data ownership may exist, and it is important that the data access system deals with this appropriately. One specific example of this is where derived data (which may be the only data available) exist only in the form of images published in a journal, with associated copyright issues. It does not at present appear that matters of data ownership will demand TCS functionalities, but appropriate ICS functions will be necessary.

Some TCSs also envisage offering specific "services" to users. The DSS community should discuss what functionalities it might like to be available, and if these can be part of the ICS or if they would require a formally established TCS. Some possible functions are further discussed below.

For some types of data, some possible or necessary TCS functions are easily identified. In other types of data this may be more difficult. There may be many different types of function which a TCS might offer. In order to effectively choose which functions it is most necessary or desirable to develop, it is beneficial if a TCS has a scientific (as opposed to purely technical) vision of what the TCS is intended to achieve. It follows that it is important that at an early stage a potential TCS develops and documents a scientific vision, and that this vision is reviewed and updated as the science develops. It can even be argued that good scientific reasoning and clearly defined scientific objectives linked to data are fundamentally necessary for effective design of any TCS structure.

2.4 Some possible DSS functionalities within, or attached to, EPOS

DSS data must be discoverable via the EPOS ICS. The natural structure for this should naturally include information on individual projects. As mentioned above, the most natural model for importing metadata information appears to be that individual data owners actively import the information. EPOS should have an effective system for this.

"DSS" data exists in different forms: Different forms of raw data, and different forms of derived data. While the form of metadata for raw data and derived data may be very different, the DSS metadata definitions need to include everything from the raw data to the various DSS "products" (derived data), including e.g. velocity models and geological interpretations of these, and even rheological and deformation-history models. Clearly, it is important that the metadata definitions for the different levels of processing are transparent and consistent with each other i.e. that there are clearly defined formalized forms of description of e.g. a derived geological model, at some suitable level of detail allowing users of the model to assess correctly how the model may be interpreted.

The SERA working group has already established a small, incomplete, database of DSS projects. Because it must ultimately be data owning institutions who provide metadata and data to the EPOS system, it is not self-apparent that it is an appropriate ambition with SERA project to further develop this database. However, in discussions with the community it has become apparent that it is advisable to do this as far as reasonably possible, not least because currently existing data from some institutions may be lost completely if proactive action is not taken soon in order to identify where the data is and thus encourage that it be secured for the future. The SERA working group will continue to request information from the community regarding such data.

Given that some institutions have difficulty in securing information on existing media, a TCS-type function could be to offer support in reading data from outdated types of media, high resolution scanning large paper sections, etc. As such an operation would presumably be distinctly time-limited, this has perhaps more of the character of a "project" as opposed to a continuous TCS-type operation. Simply supplying information regarding where facilities for reading outdated media still exist could largely solve the tape-reading problem. As many libraries etc are now undertaking large scale

digitization of books and other material in their collections, there are probably many facilities in Europe with the technical capability to scan large images at high resolution, but where these are and how access might be obtained may not be information easily available to the DSS community. A supply of information on such matters could facilitate digitization of the relevant material.

Various data presentation functions will be necessary. Some of these, e.g. producing maps showing DSS profile locations, must reasonably be part of the ICS functions.

It seems natural that EPOS should offer support in gaining access to software for analysis of DSS data. In its simplest possible form, this could simply be information on the existence of relevant software, and who to contact to request access to the codes. As access to software will be an issue for many parts of EPOS, having these functionalities within the ICS would appear to be appropriate. One issue which was mentioned in discussions with the DSS community is that some relevant processing is achieved using freely available academic software, but some using commercial software. This implies that some functionalities may be desirable to e.g. help interface between commercial and academic software. It could also be that a TCS function offering access ("transnational access") to commercial software for users from institutions who do not have such access may be useful.

EPOS, and/or the associated TCSs, are likely to offer some data analysis functions. One possibility is that processing is done by the EPOS function, another is that the system offers very convenient downloading of software and data, allowing the processing to be done on the user's computer. What types of data processing would the DSS community like EPOS to offer? It would e.g. seem attractive if it was possible to define filters and plot sections, as part of initial investigation of a given data set. Some such processing operations may reasonably be achieved using the ICS. More detailed or advanced functions could require a TCS.

EPOS must offer appropriate tools for data tracking, allocation of credit for data owning institutions, limiting data access as appropriate, etc. In general, such functions must be part of the ICS system. This may be of central importance for DSS data, as some of this is owned commercially. Even data collected on a non-commercial basis may be commercially interesting, and some data owners may be interested in selling access to data. The EPOS system must be able to deal with such matters appropriately.

Some TCSs envisage "soft" functions of more administrative nature. For DSS data, this could involve e.g. some kind of support (primarily via information flow) regarding planning of DSS projects, possible availability of DSS equipment etc. It is not immediately obvious that a TSC function for this is motivated. However, some structured access to information, e.g. regarding existing DSS-relevant instrumentation at different institutions, could perhaps be a function within the ICS (most such instrumentation is also relevant for non-DSS studies)

If the DSS community does wish to have some kind of support function related to project planning, then this is likely best achieved without a dedicated TCS. One reason for this that the seismology TCS already has a register of seismological instrumentation, much of which is relevant both for DSS and other types of data acquisition. Some types of instrumentation may differ for DSS and some other seismological measurements (types of sensor, sensor strings, frequency response….), but considering including even information on all such equipment in a common database would appear natural. In general, the distinction between "passive" and "active" source seismology has recently become rather diffuse, e.g. with passive methods now being used routinely together with active measurements. There is currently no component in EPOS relating to available seismic sources, and to issues related to these (e.g. formal limits on what may be allowed to be done in particular places, permitting, etc). Some simple channel within EPOS for information exchange on such matters might be appropriate, but this would not appear to be a major undertaking.

Especially when reflection data is included, very large amounts of "DSS" data exists. Especially because this data may be different in character, e.g. regarding the source receiver geometries, there are many different possible forms of processing and analysis. Facilitating access to data in a homogeneous

manner should also open for the development of new processing methodologies, e.g. combining data of rather different character or geometry. This suggests that development projects after SERA may be well-motivated. SERA will discuss this with the community, with the intention of facilitating discussions about one or more further pan-European projects to further develop tools for DSS analysis. A first step here might be to try to identify which new functions the community is especially keen to develop.

2.5 Are TCS functions necessary for DSS data, and if so is a new TCS necessary, or could one of the existing TCS functions be expanded?

Irrespective of organization, metadata formats must be developed and agreed upon. Some expert knowledge of DSS data is necessary for this to be achieved successfully. Coordination of this process is necessary. This need does not necessary imply that a new TCS unit is needed. While some updating of metadata definitions will doubtless sometimes be necessary, it would appear that the major job with metadata definitions will be an initial, time-limited phase leading to definition of the metadata structure. Irrespective of if there is a DSS TCS, intimate coordination with the EPOS ICS and probably also other TCSs will be necessary in order to achieve full functionality of the EPOS system.

There is no strict definition of what depth range within the Earth the term "DSS" refers to. However, generally speaking most of the relevant data elucidates the upper few tens of kilometres of the Earth's crust and mantle, down to depths of perhaps 50km or 70km. This means that the DSS data can reveal primarily structures with the crust. Clearly, this relates to surface and near-surface geology, including e.g. information from boreholes. Interpretation of DSS data is facilitated by consideration of surface geology and understanding the surface geology is aided by the DSS data. This could imply that the DSS data might best be dealt with in an integrated manner by inclusion in the responsibility of the TCS called "Geological Information and Modeling". However, the raw DSS data is seismic recordings of ground vibrations, i.e. the type of data dealt with by the seismology TCS. Furthermore, both the volcano observatories and the near fault observatories TCSs should include seismic data.

Organizational structures demand resources to design and operate. Finding financing for the long-term operational cost of such structures may be difficult. In addition, the underlying concept of EPOS is the integration of data. There is a risk that a large number of sub-units (such as TCSs) could cause unnecessary hinders in achieving optimal data integration. Therefore, there are some strong arguments for limiting the number of such sub-units (TCSs), possible to the extent of reducing from the number currently envisaged.

All-in-all, there appear to be only relatively weak arguments for considering a new TCS specifically for DSS data. However, there appear to be strong arguments for considerable investments in the development of new analysis tools, which can later be integrated into, or made available by, EPOS. This could be achieved by one or more new time-limited development projects, with broad engagement from the DSS community.

One possibility is that the seismology TCS be requested to take responsibility for the DSS raw data, primarily in the sense of defining the relevant metadata and data formats. DSS data products, especially those where complementary data (e.g. geological information) has been used in analysis leading to the product, do not appear to fit well into the seismology TCS, but rather that focusing on geology. Thus, it would seem more appropriate that the "Geological Information and Modeling" TCS was responsible for this higher level derived data. It is not immediately clear at what level responsibility should move from the seismology TCS, but this does not appear to be a major issue.

3 Current proposal for the continued process

The DSS community sees a need for a well-functioning DSS component within EPOS.

This should from the beginning include both “wide angle” and “reflection” data.

A trial system for this should be developed, within SERA.

Import of metadata and data access information to EPOS should be passive i.e. the data suppliers import the data, rather than EPOS or some active function offered by a TCS (which would likely be inappropriately expensive)

No dedicated TCS for DSS data appears to be motivated

The seismology TCS should take responsibility for raw data. This would appear to be a matter demanding only rather limited resources, because of the considerable overlap with seismic data in general.

The TCS “Geological Information and Modeling” should be responsible for the more advanced derived data. Metadata structures must be carefully designed such that the DSS products can be easily and robustly used together with other data on e.g. crustal structure.

A dialogue with colleagues planning the ICS and with the relevant TCSs will be necessary to coordinate further planning of how DSS data will best be dealt with.

With the help of the DSS community, the SERA working group should continue to refine the DSS project database, with a primary short term aim of securing data which may be in danger of being lost.

For DSS data, and especially reflection seismic data, there may be commercial interests involved. It is completely necessary for the EPOS ICS to offer full security in maintaining control of the use of data, if data owners are to be prepared to offer data access via EPOS.

A software repository function for DSS is desirable. This is probably best achieved as part of a consistent such function within the EPOS ICS. There do not appear to be strong motivations for EPOS to offer major processing facilities for DSS. However, some basic functions for e.g. display and filtering will be very helpful. These can likely be rather generic in character, not specific for DSS.

An information service regarding securing data stored on outdated media is desirable and should be achieved short-term.

The possibility of offering “trans-national access” for access to advanced processing tools should be investigated.

There appear to be strong motivations and interest for future development projects in order to develop new tools for DSS data analysis. The SERA working group will continue to discuss possibilities with the DSS community.

Dialogue with the DSS community will continue. The next structured initiative will be at the Deep Seismics meeting in June 2018.

Contact

Project lead	ETH Zürich
Project coordinator	Prof. Dr. Domenico Giardini
Project manager	Dr. Kauzar Saleh
Project office	ETH Department of Earth Sciences
	Sonneggstrasse 5, NO H-floor, CH-8092 Zürich
	sera_office@erdw.ethz.ch
	+41 44 632 9690
Project website	www.sera-eu.org