Annexe 1: Learnings from the Fakara case study, and recommendations for research data documentation at ICRISAT

Let us briefly review the major incentives of metadata creation. They can be summarized in six categories: i/ help potential users *retrieve data* and *evaluate fitness*, ii/ help data producers *publicize and support use* of data, iii/ *increase the value* of data as potential users are more likely to retrieve information about it and make proper use of it, iv/ *protect an organization's investment* in data throughout the years, v/ *limit loss of value* that affects undocumented data with staff changes, and vi/ *reduce duplication* of datasets arising from lack of confidence in existing data.

The various advantages associated with the efficient and effective production of relevant metadata are hardly disputable: proper research data documentation is very important and an enabling environment is required. However, the potential high burden of metadata creation calls for special attention when devising dedicated institutional mechanisms. Learnings from the Fakara case study have been compiled below with synthetic recommendations for data documentation at ICRISAT:

1. Dedicated human resources are mandatory. At the technical level, metadata creation cannot be done without encoders, who play a role comparable to genebank technicians or librarians. They sort, clean and store metadata records and maintain the integrity and security of the metadatabase under the supervision of a data manager (equivalent of the chief librarian). The complex nature of metadata edition requires dedicated time which is not available in most scientists schedules, and specialized skills which are seldom found in many research assistants.

2. Resources should be shared, but tied to projects. There is a danger in creating 'datacratic' positions which would be disjoint from project needs and objectives. One reasonable option could be to hire one data manager per region (ESA, SEA, WCA) for proximal coordination and backstopping. Local encoding skills would be developed at the country level, either through one IT and/or GIS technician availed part-time to a suite of projects, or through capacity building within existing project staff. Oversight and commitment of project leaders should be sought.

3. Raising awareness among scientists is essential. Of particular importance is the need to build trust, by explaining that i/ sharing metadata is not about releasing one's data in the wild, and that ii/ appropriate restrictions can be easily controlled by scientists for adequate data security. Building trust in the process of data documentation will also be achieved by sensible use of scientists' limited time. This in turn requires good interpersonal skills in metadata encoders in addition to their technical capacity.

4. Software solutions are generally not a constraint, but they vary in complexity and across scientific disciplines as do metadata standards and formats. There is no one-size-fits-all metadata editor or utility, which substantiates the need for dedicated, conversant staff. Some software (e.g. M3Cat) allow for quick learning and direct use by non-specialists (e.g. project leaders). Many are network-based, accessible through web browser interfaces, and open-source, hence easing procurement, deployment and scientists' input.

5. Targeted investments can efficiently document past data. Altough *a posteriori* metadata encoding requires additional efforts (resources), limited investments can go a long way when areas of interest have been identified by donor partners. One successful approach to priority setting is to pinpoint geographical areas of project overlap, as in the Fakara region. A list of similar benchmark sites (Kenya: Machakos, Zimbabwe: Tsholotsho, etc.) can be assembled and showcased to potential donors as low risk, high return investments for past data salvage – especially when built in project proposals by scientists.

6. Reliable network connectivity is important when working with a large group of data producers in a decentralized structure. Network-based tools, open-source or commercial, can significantly decrease the time and costs of data documentation, (meta-)database synchronization, versioning, and internal consistency. The Fakara exercise has demonstrated that connectivity in ICRISAT-Sadoré (Niger) is not adequate with insufficient bandwidths. Other ICRISAT locations probably face similar constraints (e.g. ICRISAT-Matopos, Zimbabwe). Close interactions with IT personnel is critical for the successful implementation of distributed (meta-) databases.

7. A (meta-) data management policy is needed. *Inter alia*, it should define ICRISAT's data lifecycles (data sharing timeline), obligations for data producers from a data documentation perspective (building metadata creation within projects, metadata sharing timeline, compliance with accepted standards and formats, etc.). It should emphasize the need (obligation?) to plan for metadatabase creation upfront at the time of project inception. In addition to future datasets, it should also cater for past data, which is by far the biggest burden facing an organization as many data creators have left.

8. A visioning workshop on data management is recommended. It should be trans-disciplinary and involve data-intensive and less intensive groups; field and laboratory data producers; genetic (e.g. bioinformatics) and environmental (e.g. GIS) groups; IT, library services, management. It should not focus on the definition of minimum metadatasets (done in the 1990s), should marginally address the issue of metadata standards/contents (mostly for information purposes), and should mainly concentrate on finalizing an Institute-wide (meta-) data policy with enabling/enforcing mechanisms: resources, rules and tools to facilitate (meta-) data flow.

9. A technical (meta-) data management task force is advisable. It would strive to facilitate the exchange of information and software solutions to customize and automate the process of research data documentation. It should foster an enhanced level of interactions between IT staff and the different research teams, that reaches beyond traditional hardware and networking issues to address specific programmation and computing needs: interfacing software from different disciplines, improving the basic ‘batching and scripting’ ability of research staff, etc. in pursuit of enhanced, coordinated institute-wide data management.

10. Online serving of (meta-) data is the ultimate goal. However efficient and effective the sharing of information is within a group, a project team, or the Institute, the largest benefits of (meta-) data creation are reaped when the latter is published before a wider audience of existing and potential partners (with appropriate security restrictions, e.g. through granularity). In the CGIAR terminology, the concept of (meta-) data serving is intimately tied to that of International Public Goods (IPGs). It is important to realize that documenting existing and past data can prove a cost and time effective way of posting IPGs. To that purpose, more attention can be directed to developing resources such as the ICT-KM program and associated tools, such as the CSI-sponsored GeoNetwork (an FAO-born open-source solution for networked, georeferenced (meta-) data serving). The Fakara metadatabase will need to be visible shortly on the ICRISAT GeoNetwork node.

11. Georeferencing field data should be mandatory. The value of numerous field data (trials, experiments) can significantly decrease when their spatial location is not adequately consigned. While there are ways to georeference ground data *a posteriori* (e.g. using village names and gazetteers), the recovery process almost always involves some loss of precision and usability. In the era of cheap GPS, GPS-patched PDAs and other navigational gadgets it is not acceptable to gather ground data without geographical coordinates. There are many electronic data collection tools to plan and facilitate the collection process, some better than others. **Paper survey sheets should be a thing of the past**. Advanced expertise in the design and use of electronic data forms with GPS-enabled field computers is available from ICRISAT GIS staff, along with high-precision georeferencing solutions for field-scale processes.