



Vladimir Alekseyev

NCBI Short Read Archive Structure

July 27, 2007

Abstract

In early 2001 NCBI, in collaboration with Ensembl, developed the Trace Archive. This has successfully served as a repository for the chromatograms produced as part of traditional Sanger based sequencing protocols for many years. In recent months, advances in sequencing technology have produced what is only the first wave of next generation sequencing technologies (e.g. 454, Illumina, ABI Solid, Helicos). Due to the structure and volume of this data it is clear that it does not efficiently and effectively fit in the current Trace Archive design, so we have decided to build a new archive for such data, the Short Read Archive (SRA).

The SRA project is well underway and is being built in collaboration with Ensembl, sequencing centers and the vendors themselves. The deployment of the SRA database is anticipated in Fall of 2007.

In the transition period, we will still accept submissions of new generation sequencing projects via secure FTP. We will hold these projects and provide access to them on a per project basis via anonymous SRA FTP. When the SRA is fully functional, all such submissions will be loaded. We will post notifications and progress updates on Short Read Archive web site.

As some of you are aware, the current Trace Archive was modified to accommodate 454 reads and a small number of such reads have been deposited there. However, this modification is only capable of handling a relatively small number of 454 reads for small projects during this transitional time. It was never expected to accommodate large volumes of 454 data (and doesn't handle the other new technologies at all). So most large 454 submissions to Trace will be held in the temporary SRA FTP site, and moved into the SRA database in the Fall.

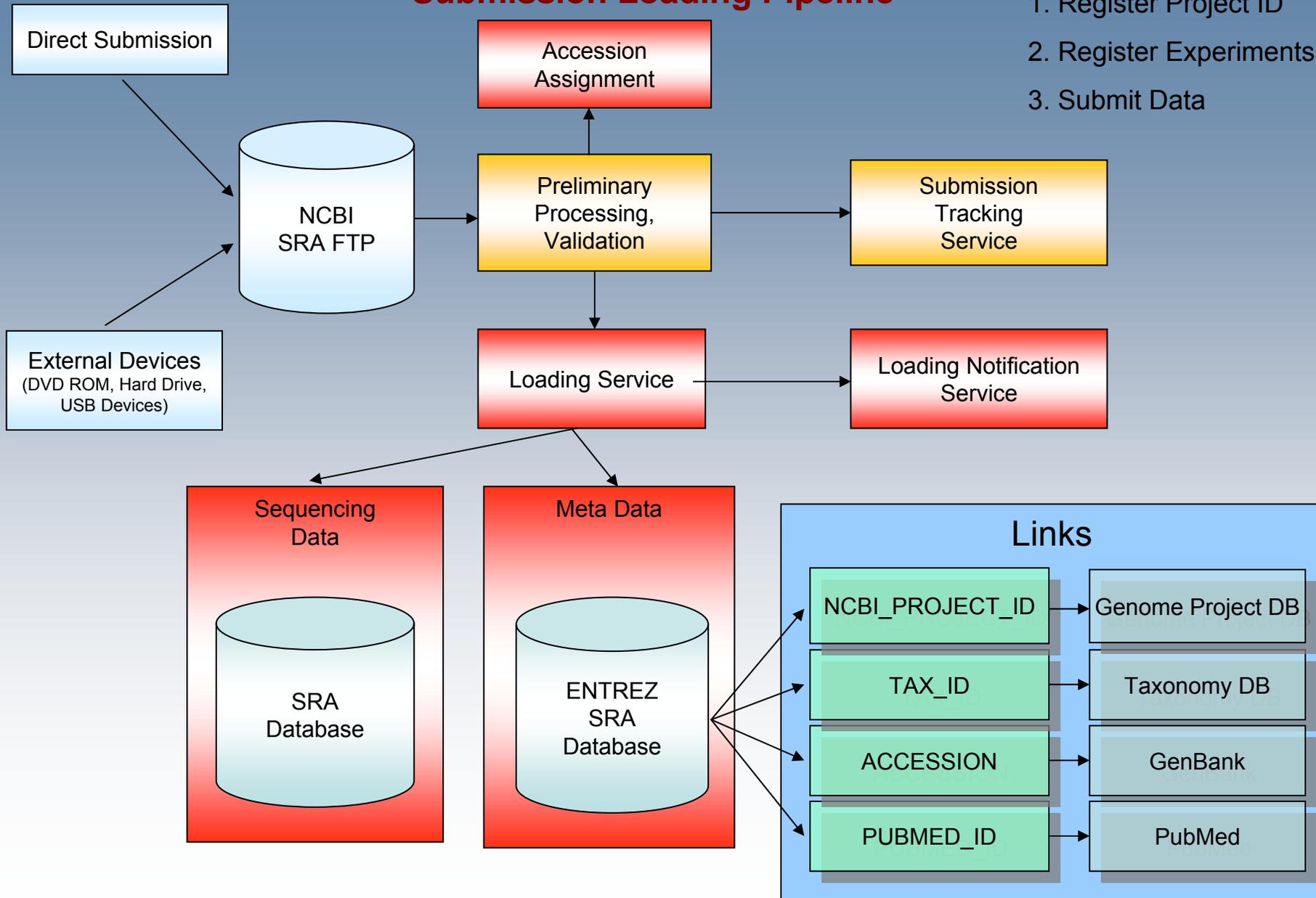
**Scope of NCBI Short Read Archive**

1. SRA is a repository for short read experiments at the level of primary base calls, qualities, and intensities. We have attempted to create a common representation for the sequencing data produced by 454/Solexa/ABI SOLiD/Helicos platforms.
2. Unit of submission is an experiment. The meta data is specified on the level of the experiment, not individual reads.
3. SRA is not intended as a place to store secondary analysis such as assemblies, alignments, or oligo profiles. Other resources will be accepting the secondary analysis data.
4. SRA will be fully integrated with the Entrez system, which allows data aggregation through the common fields and linking to other resources.
5. SRA is not a place to store raw image data (much too large to be practical).

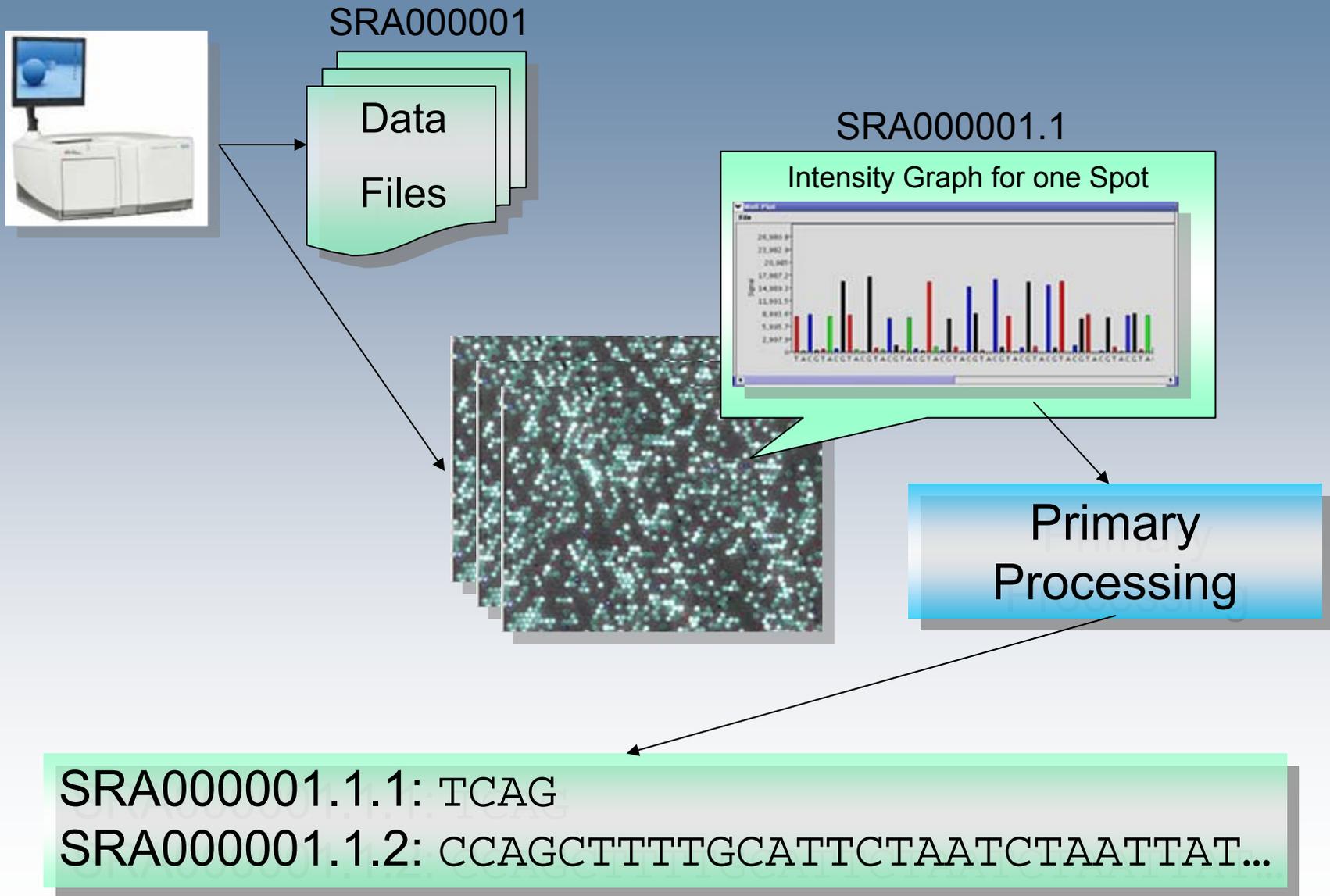
NCBI Short Read Archive Structure

Submission Loading Pipeline

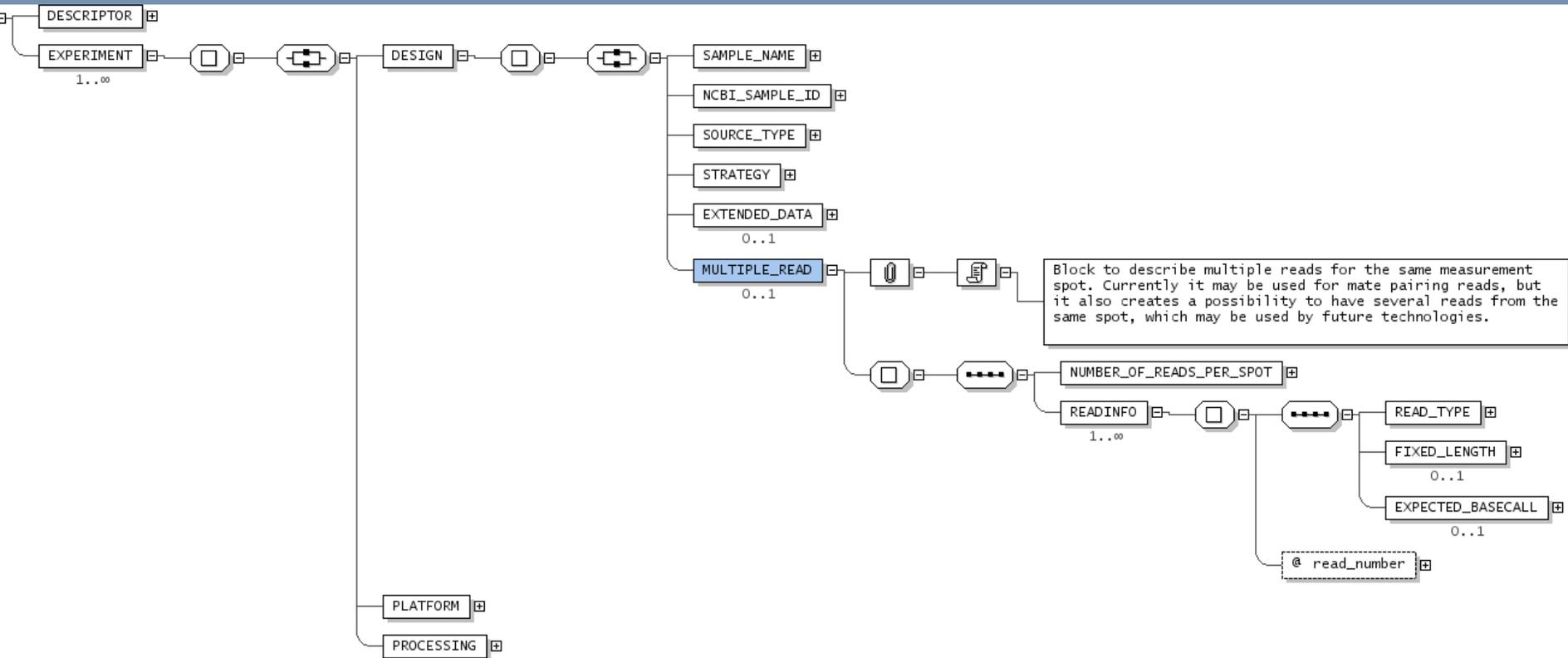
1. Register Project ID
2. Register Experiments
3. Submit Data



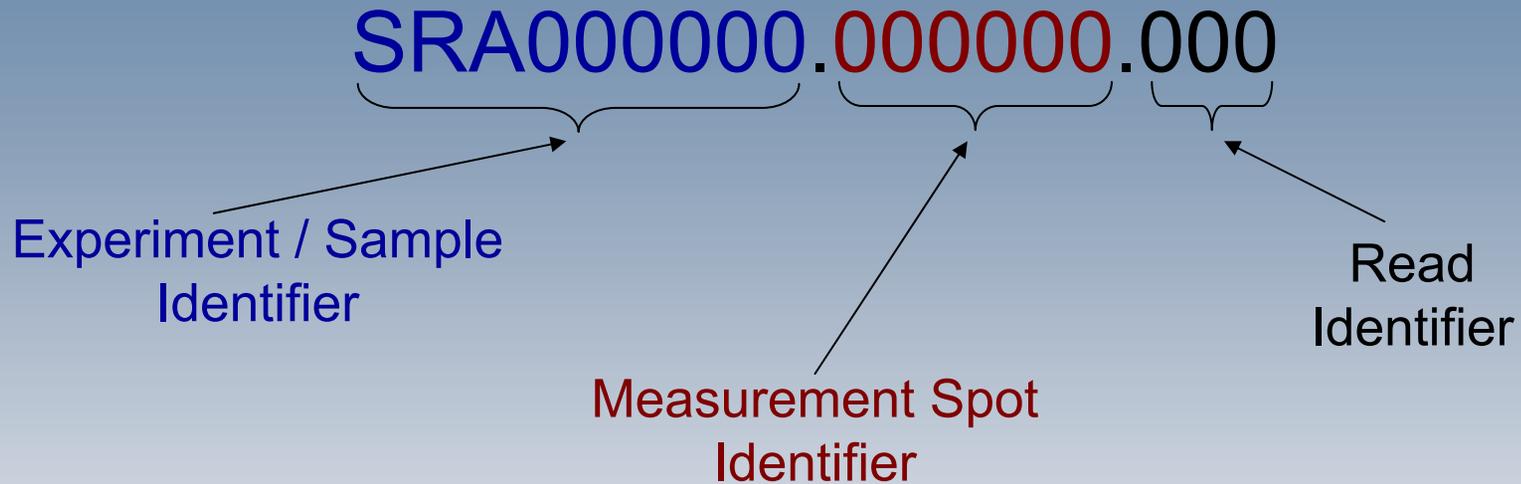
Experiment Data Structure



Meta Data: Multiple Read



Accessioning proposal

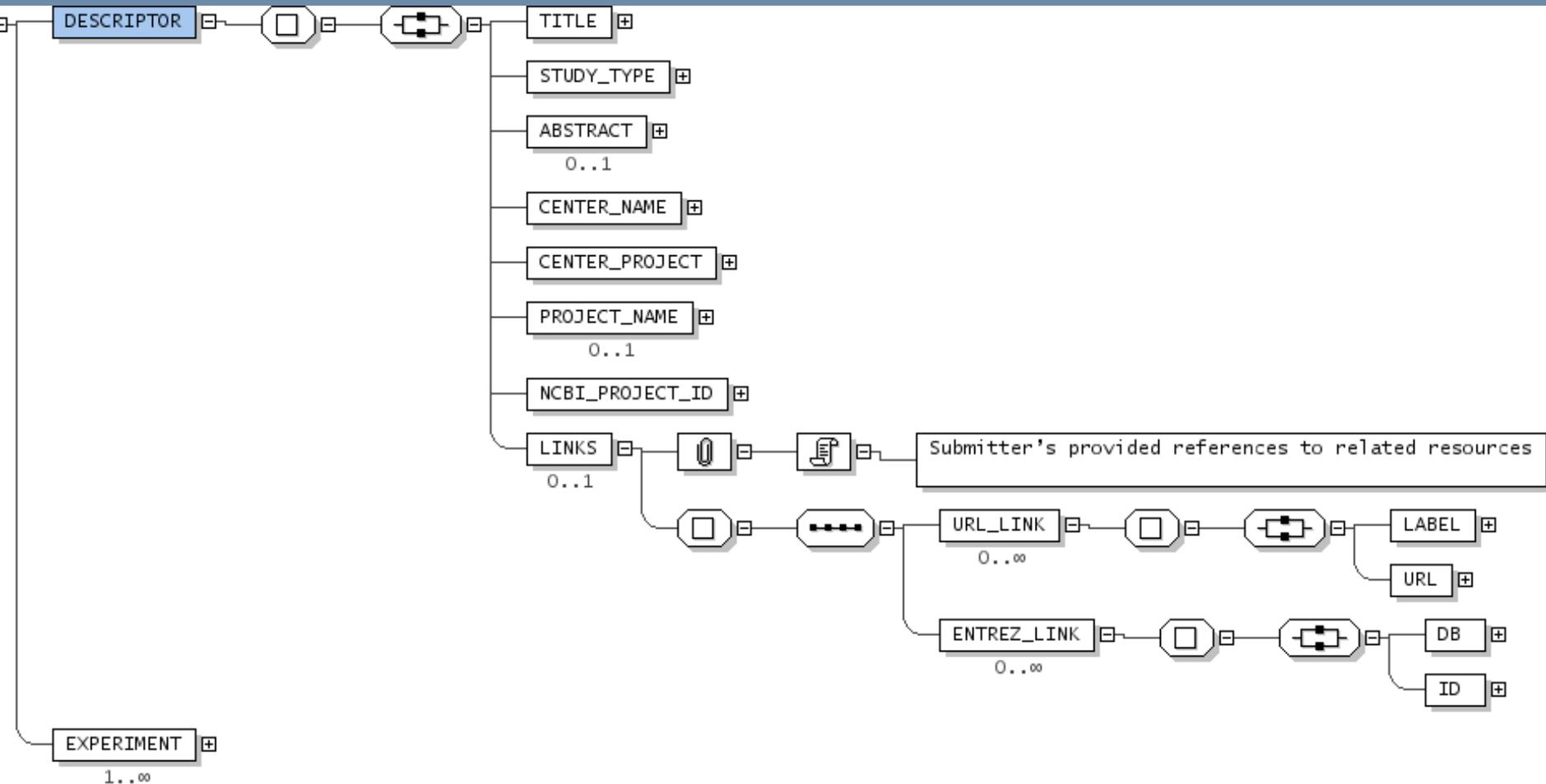


Examples:

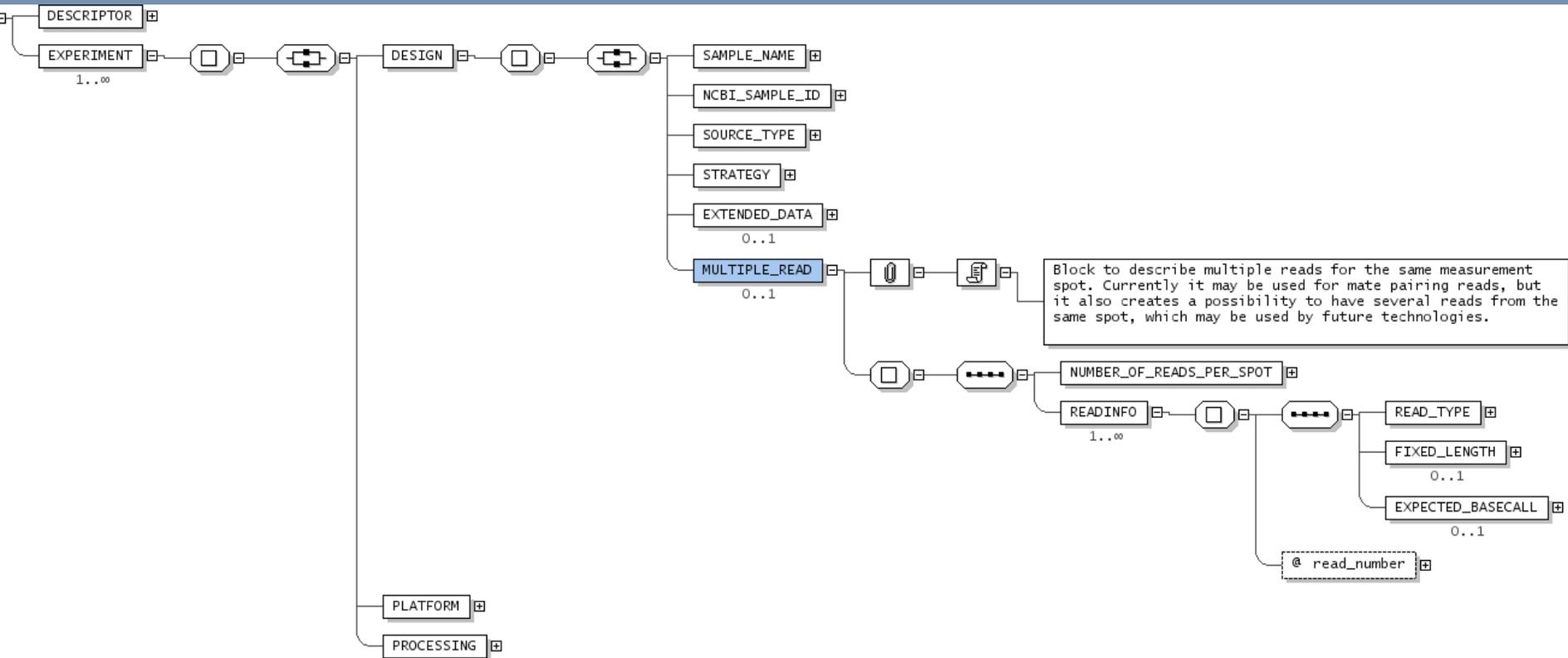
- | | |
|---------------|--|
| SRA000001 | Addresses metadata and the whole experiment |
| SRA000001.1 | Addresses intensities and all reads from the spot |
| SRA000001.1.1 | Addresses individual reads. Will be used in secondary processing |

NCBI Short Read Archive Structure

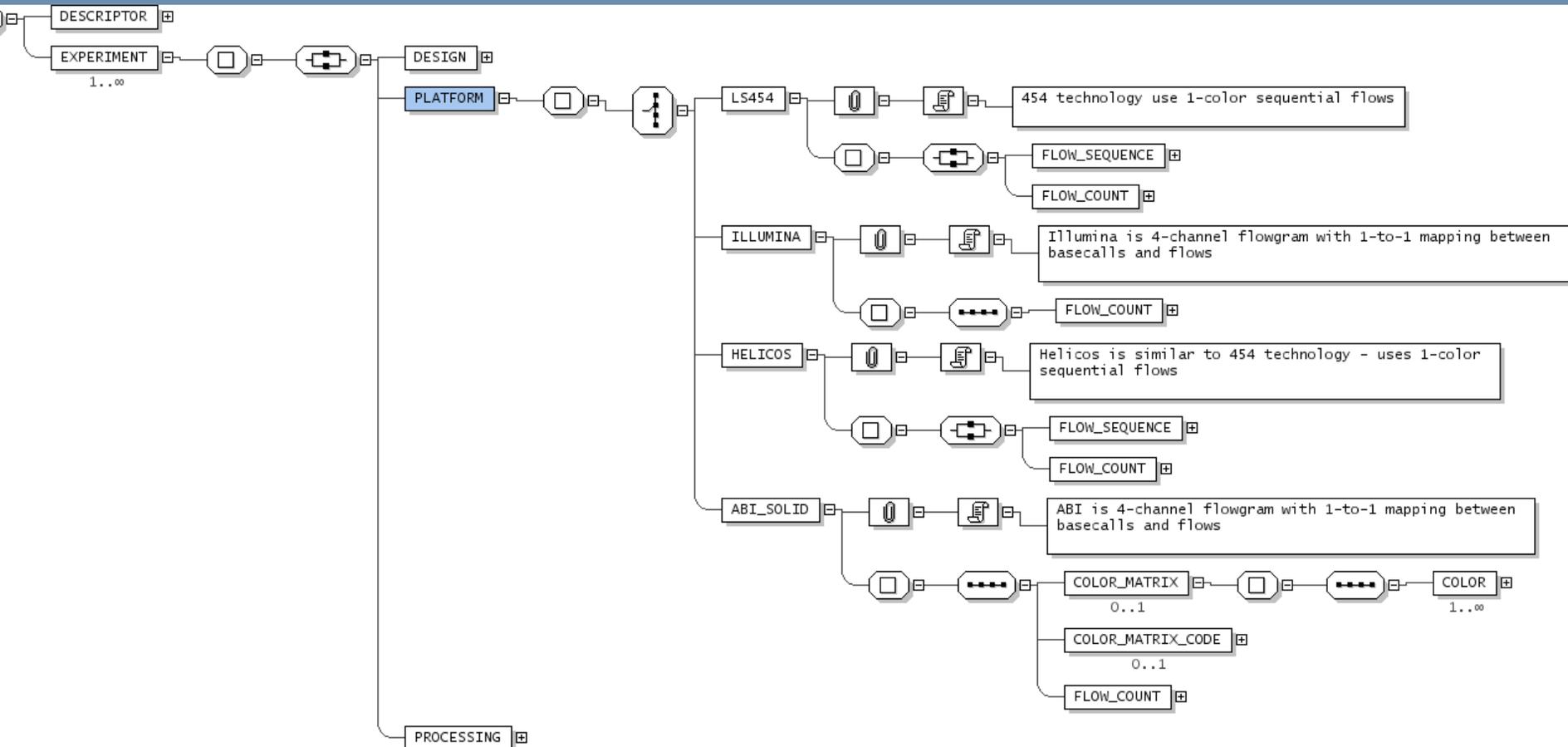
Meta Data. Descriptor



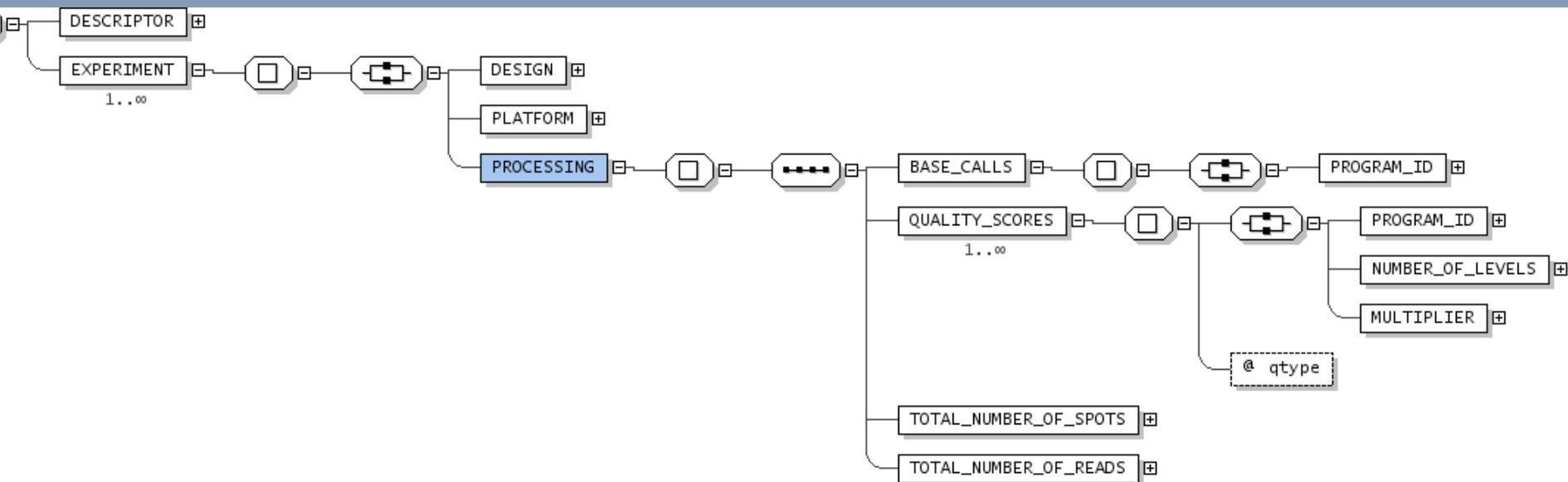
Meta Data: Multiple Read



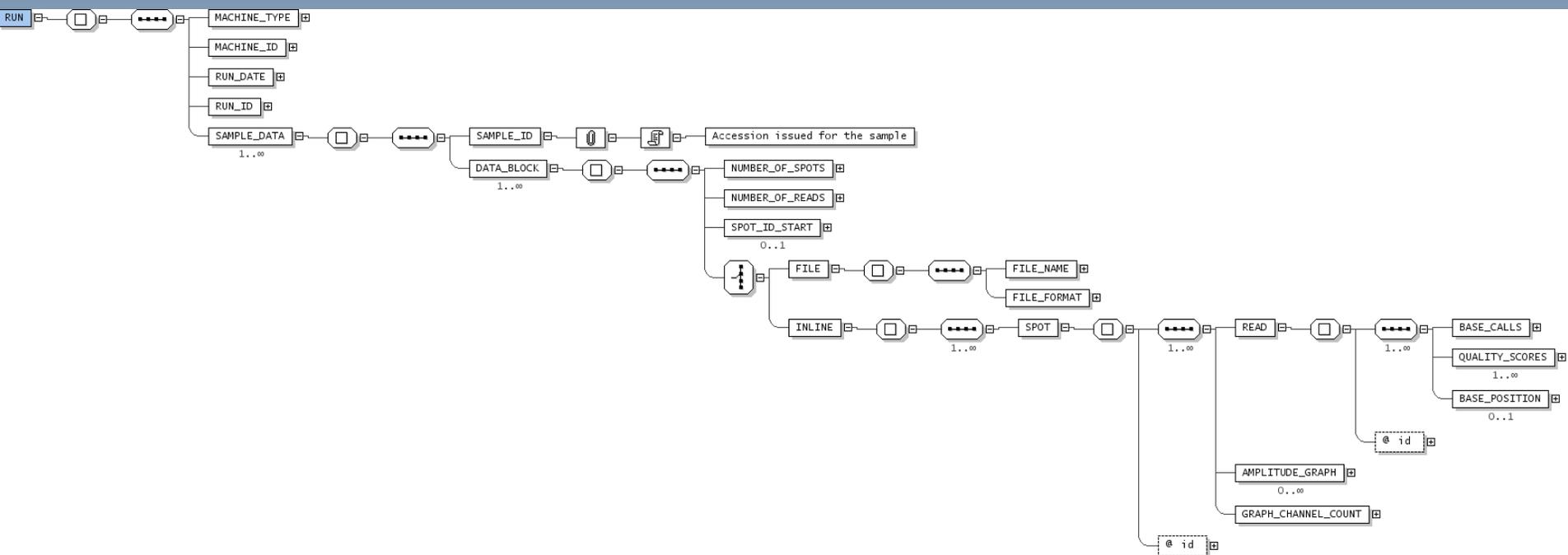
Meta Data: Platform



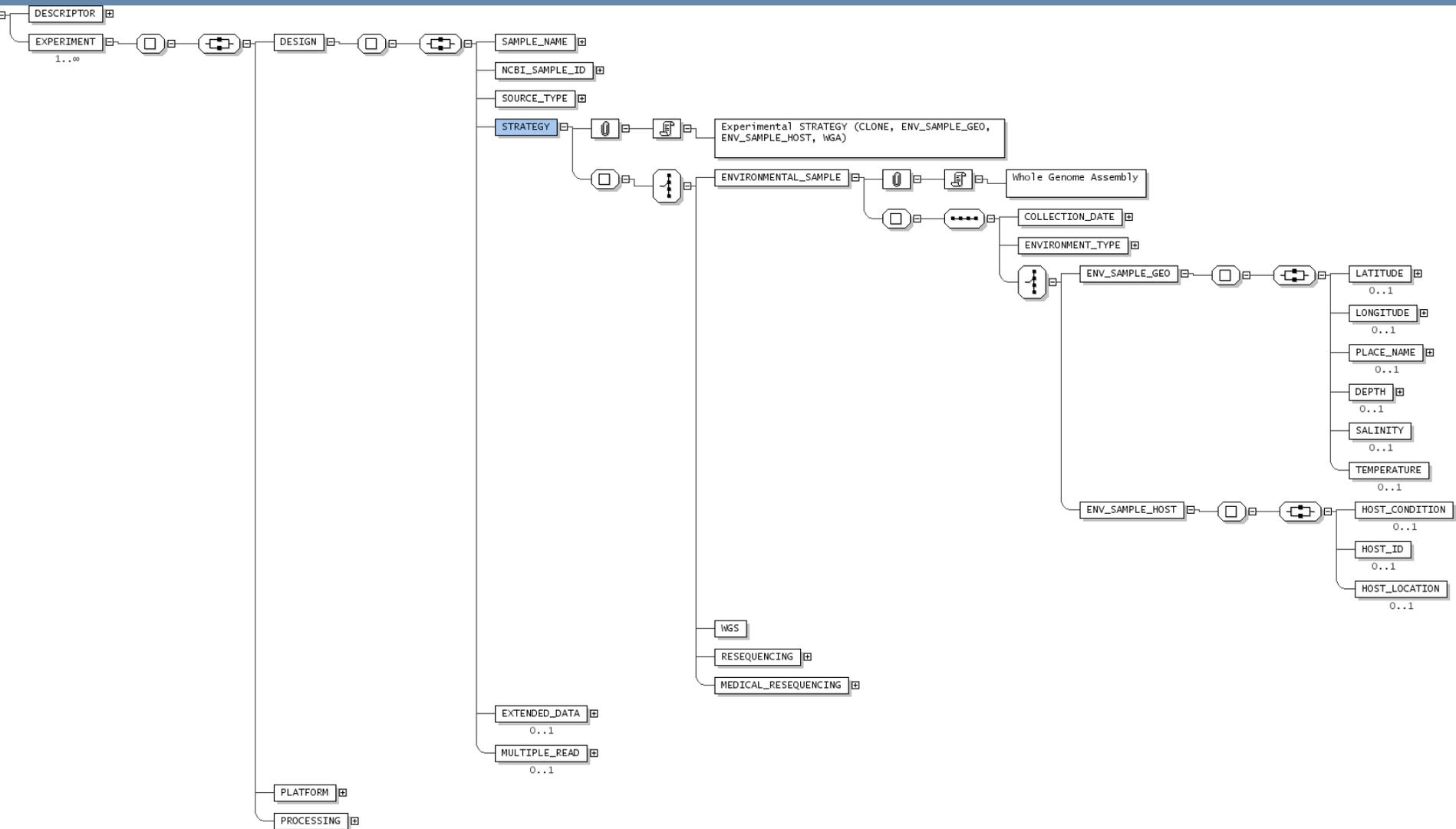
Meta Data: Processing



Meta Data: Run



Meta Data: Strategy



Acknowledgments

Eugene Yaschenko

Deanna Church

Alexey Egorov

Sergey Ponomarev

Alexey Vysokolov

Kurt Rodarmer

Dmitry Volodin

Martin Shumway

Vladimir Alekseyev
aleksey@ncbi.nlm.nih.gov

NCBI Short Read Archive Structure

July 27, 2007