GEOLOGICAL SURVEY OF CANADA

OPEN FILE 7509

Structure and Data Quality Assessment of the Kimberlite Indicator and Diamond Database (KIDD)

J.-E. Lesemann, C. Fuzz, B.A. Kjarsgaard, and H.A.J. Russell

2014

Natural Resources Canada | Ressources naturelles Canada

Canada

GEOLOGICAL SURVEY OF CANADA

OPEN FILE 7509

Structure and Data Quality Assessment of the Kimberlite Indicator and Diamond Database (KIDD)

J.-E. Lesemann[1], C. Fuzz[2], B.A. Kjarsgaard[3], and H.A.J. Russell[3]

[1] Vancouver Island University, Nanaimo, British Columbia

[2] University of Western Ontario, London, Ontario

[3] Geological Survey of Canada, Ottawa, Ontario

2014

doi:10.4095/293334

This publication is available for free download through GEOSCAN (http://geoscan.ess.nrcan.gc.ca/)

Recommended citation

Lesemann, J.-E., Fuzz, C., Kjarsgaard, B.A., and Russell, H.A.J., 2014. Structure and Data Quality Assessment of the Kimberlite Indicator and Diamond Database (KIDD); Geological Survey of Canada, Open File 7509, 33 p. doi:10.4095/293334

Publications in this series have not been edited; they are released as submitted by the authors.

Abstract

The Kimberlite Indicator Diamond Database (KIDD) developed by Northwest Territories Geoscience Office is a relational database archiving kimberlite indicator mineral (KIM) grain counts reported within assessment reports of mining activity in the Northwest Territories and Nunavut. KIDD archives four main types of sample data and metadata: 1) sample and site descriptions (general sample attributes: sample number, location, etc.); 2) information on KIM (grain count data for suites of KIM); 3) analytical information (processing techniques, processed size fractions, etc.); 4) comments (mixed array of sample site attributes, analytical information, KIM information). The completeness and accuracy of reported KIM grain counts is variable. There are KIM grain count entries for only ~34% of all archived samples. Entries for individual KIM grain counts vary between ~0.02% (diamonds) to ~16% (garnets).

Despite some limitations in grain count reporting, the data contained within KIDD are of high quality and integrity: 87-100% of data reported within assessment reports are faithfully and accurately reported within KIDD. Limitations to the use of KIDD also result from irregular and non-standardized inconsistent reporting of sample weights and sampling site attributes. These limitations result not from an absence of data but from the database structure itself that does not contain proper fields for standardized reporting of these attributes.

Overall, KIDD offers a useable archival dataset of high data quality. However, numerous caveats and limitations in data reporting require careful data evaluation by the user.

Table of Contents

1. Introduction

Canadian provincial and territorial mining regulations require exploration companies to submit Assessment Reports of work carried out on staked mineral claims to maintain their status (e.g. Northwest Territories and Nunavut Mining Regulations, 2013). These Assessment Reports, though publicly available, during the 20th century were largely submitted in print form to regional geoscience offices, which limited their public availability and usage. More recently these reports can be submitted in both print and digital format, which is part of a global trend (e.g. NSW Industry and Investment, 2011). The kimberlite indicator diamond database (KIDD) was developed in 1999 at the Northwest Territories Geoscience Office as a means of compiling and centralizing datasets contained within assessment reports (Armstrong and Lee, 2000). This initiative was an early effort to facilitate dissemination of exploration data (sample locations and results) that was only available in print. A rationale for development of KIDD was to ensure easy access to Crown information during preliminary stages of economic development by facilitating prioritization of exploration targets and ground selection (Armstrong and Lee, 2000). Since 1999, the KIDD has expanded as ongoing diamond exploration in northern Canada led to a proliferation of assessment reports. Additions to the database since 1999 have increased the potential interest and value of this archive of exploration data. Indeed, KIDD has been heralded as a valuable exploration tool for many stakeholders in the exploration industry (Armstrong, 2003; Paulen, 2012; Jones, 2013). Despite the nearly 15 year history of KIDD there has been limited documentation of the database structure, and no formal assessment of its organization and content. Armstrong and Lee (2000) warned against potential 'introduced caveats' (erroneous entries) associated with digitizing errors, character recognition errors, or summation errors. At the time, no attempt was made to identify and rectify these entries. The potential frequency and significance of erroneous entries is unknown, and their effects on data quality and integrity within KIDD have yet to be assessed.

1.1. Global Mineral Exploration and Public Domain Data Framework

It is commonly acknowledged that public geoscience knowledge is one of Canada's competitive advantages in attracting mineral exploration (Duke, 2010). In this regard Canada is in competition with other major mining regions to maintain and improve the geoscience framework to attract and maximize exploration efficiency and success.

Public geoscience increases exploration efficiency in a number of ways; one is by reducing duplication through provision of common information in the public domain. KIDD is a notable component of improving the accessibility to Indicator Mineral data in the assessment reports submitted by exploration companies in NWT and Nunavut. Similar initiatives have subsequently been undertaken or are underway in other jurisdictions within Canada (e.g. Keller and Bogdan, 2004) and internationally (e.g. NSW Industry and Investment, 2010; Australia Victoria Mineral Exploration Geochemistry Data; Sweden Exploration Reports). In fact KIDD is simply a recent example of decades of government and industry attempting to maximize exploration success through archiving and sustaining accessibility to both industry and government collected datasets (e.g. core repositories, Simpson, 1985).

Other fields of geoscience pursue similar objectives of private sector data collation for publication for the public good. Notably, water well drilling records are commonly required to be submitted to the appropriate government agency across Canada (e.g. Ministry Ontario Environment, 2013). Common with mineral, and petroleum data is the issue of common reporting formats (e.g. water well records, Russell et al 1998; petroleum data, PPDM Association, 2013). The issue of data access and standardization was recently a focus of a Geomapping for Energy and Minerals (GEM) project on the compilation and web enabling of industry data (Paulen, 2012). As the geomatics industry recognizes the opportunity for value added distribution and analysis (Jones, 2008) documentation of the data structure and data reliability becomes increasing important. The improvements to KIDD within the geoscience data structure of the NWT can continue to advance needed improvements to enhance support for mineral exploration (e.g. LookNorth 2012, p. 14).

2. Objectives

This report has three objectives:

1) Describe the overall KIDD structure and data organization scheme.

2) Assess the KIDD content in order to evaluate its usefulness.

3) Suggest ways to maximize its potential as an exploration tool and its ability to fulfill its stated mandate of facilitating access to Crown information.

3. History of the KIDD and affiliated user interface

KIDD is a relational geospatial database currently accessible online via a graphic user interface maintained by the Northwest Territory Geoscience Office (NTGO) (http://ntgomap.nwtgeoscience.ca/). KIDD forms part of the *Diamond Database*, a larger data archive of diamond exploration data in the Northwest Territories that also contains the Kimberlite Indicator Mineral Chemistry data (KIMC). KIMC data includes mineral chemistry information from picked kimberlite indicator minerals (NTGO website, 2013).

As of early 2013, KIDD contains data from reports entered up to 2011. This consists of 638 assessment reports, and two Open File reports from the Geological Survey of Canada. Together, these reports account for 219,770 distinct records within the database.

3.1. Phase 1 data entry (1999-2000)

Initial data entry at inception of the KIDD project (April 1999, Armstrong and Lee, 2000) included ~60,000-65,000 till samples from ~160 Assessment Reports. This initial data entry was produced by digitizing printed Assessment Reports, including the sample locations (as coordinates or map scans) and indicator mineral (IM) picking results (in table formats).

3.2. Phase 2 data entry (2000-2005-2011)

Up to 2005, data entry followed the established protocol of Armstrong and Lee (2000) (see 3.1 above). KIDD and KIMC data files were managed as series of individual spreadsheets (MS Excel format) (NTGO Website, 2013). After 2005 spreadsheet data were integrated into the relational *Diamond Database* where they can be queried and downloaded online (Armstrong et al. 2004).

3.3 User interface and data access

KIDD data are accessed through the 'NT GoMap' and NT GoData web portal, hosted by the Northwest Territories Geoscience Office. NT GoMap is a graphic user interface allowing query of the KIDD (and other) databases (http://ntgomap.nwtgeoscience.ca/). NT GoData provides a search interface of NWT geoscience database collection without

the mapping interface. The KIDD can be queried based on database fields listed in section 4.

Through the interface, users gain access to the KIM grain count data archived in KIDD, and can link to the original Assessment Report submissions (typically available as a scanned PDF document). Of the 640 assessment reports archived in KIDD, 68% are available for download from the NTGO website. The remaining ~32% (204 reports) are available via the Nunavut Geoscience Office. The data from these 204 reports are archived in KIDD.

3.4 Database structure and data entry protocols

Limited database structure and data entry protocols were established during development of KIDD. Armstrong and Lee (2000), however, established some clear protocols, for example null value entries in KIDD were to be defined by "-1" entries. It is currently unclear to what extent these protocols have been maintained or changed.

4. KIDD structure and data organization

KIDD archives data in tabulated form where data are arranged as a series of columns designed to contain either numeric or text inputs. Columns contain descriptive fields that can be subdivided into four subcategories (Tables 1-4):

1) *Sample and site descriptions*: contains general sample attributes, sample number, site locations, company name, etc.

2) *Kimberlite Indicator Mineral Information*: contains grain count data for suites of KIM, and total grain counts.

3) *Analytical Information*: describes processing techniques, processed size fractions, and the name(s) of companies who performed the analyses.

4) *Comments*: contains a mixed array of entries related to site attributes, analytical information, and indicator mineral information.

This report follows the existing KIDD data structure to examine and describe the characteristic of fields within KIDD. Limitations of the current structure are noted and, where pertinent, suggested improvements to the structure and data organization are made.

4.1. Sample and site descriptions: structure, content and limitations

The vast majority of samples contain a sample number (SAMPLE_NUM, Table 1), reference number (REF_NUM, Table 1), and reference type (REF_TYPE, Table 1). Each sample should therefore be traceable to an original assessment report recorded by the reference number. However, 361 samples in KIDD have no REF_NUM entry and therefore cannot be linked to an Assessment Report. This represents 0.16% of the total entries in KIDD.

The reference type (REF_TYPE) specifies the type of source containing the archived data and commonly consists of assessment reports. Out of 640 archived reports, only two are Open File Reports from the Geological Survey of Canada. The remaining 638 sources are industry-submitted assessment reports. Data are accessed via the REPORT_URL field, which links to online metadata for the specific data source, but does not link directly to the source assessment report (see 3.3 above).

Company names (COMPANY, Table 1) are not always specified in KIDD, though the information is always present on the submitted assessment report and can therefore be traced back to this source. A total of 204 unique entries (distinct company names) occur in the COMPANY field within KIDD. This field contains 63 695 null entries (~29% of entries for this field).

All samples contain geographic coordinate data (LONG and LAT fields, Table 1), allowing each samples to be rapidly plotted with GIS software. However, fields relating to site coordinates do not specify the geodetic datum used in reporting coordinates. Armstrong and Lee (2000) specified the use of the NAD 83 in their original development of KIDD and reported on the method used to convert NAD 27 coordinates to NAD83. However, these conversion details are not specified within KIDD and the limited documentation from NTGO website does not specify how the geodetic datum is treated. For example, multiple Assessment Reports within KIDD specifically refer to NAD 83 datum in some columns and it is unclear if coordinates presented under the coordinate fields follow the initial protocol outlined by Armstrong and Lee (2000) and have been converted from NAD 27, or if they are the raw reported entries from the Assessment Report under the NAD 83 datum.

Table 1: Fields associated with site and sample descriptions from Assessment Reports and other sources within KIDD.

Field	Description
SAMPLE_NUM	Sample number: a unique identifier for each sampled entered in KIDD. Multiple samples are typically associated with a single survey.
REF_NUM	Reference number: identifier assigned to a source document (commonly an assessment report)
REF_TYPE	Reference type: identifier describing the type of data source
REPORT_URL	Report uniform resource locator: link to online metadata for a data source. Does not link directly to source report, only to summary metadata sheet hosted on NTGO website.
COMPANY	Name of company that submitted the assessment report to NTGO.
LONG	Longitudinal coordinates of sample
LAT	Latitudinal coordinates of sample

Selective examination of 70 assessment reports reveals that geodetic datum information is frequently reported in these reports, yet is lacking within KIDD. Furthermore, within assessment reports some base maps used to plot data are based on NAD27 datum, while reported site coordinates were collected with GPS on a NAD83 datum. This creates a significant potential error when trying to plot and/or relocate sample sites.

These discrepancies remain unaddressed and difficult to resolve without further information and standardization of the reporting format, and uniform and consistent conversion protocols. A significant cause of the confusion can be attributed to the absence of a data field to specify geodetic datum in KIDD, despite the existence of this information in assessment reports. Additional improvements to KIDD could also include indications of bounding coordinates for each survey. This would help improve the analytical potential of KIDD by providing ways to rapidly assess survey coverage, which

can be used to estimate sample point density. Online-accessible NTGO data sheets for surveys and assessment reports provide information on bounding coordinates, although these data have not been integrated within KIDD.

4.1. Kimberlite Indicator Mineral Information: structure, content and limitations

KIDD records a suite of kimberlite indicator minerals (Table 2) and provides a total grain count for all indicators within each sample. These data consist of numeric entries. Given the main stated purpose of KIDD is to archive IM count information, and the early focus of KIDD on indicator mineral picking results, fields reporting grain counts are central to the integrity of the KIDD.

4.1.1. Data quality and accuracy assessment

The completeness and accuracy of reported grain counts in KIDD is variable. For example, examining all sample entries (n= 219 770 samples) in KIDD reveals that total grains are reported for only 74 434 samples (~34% of samples) (Table 2). Furthermore, only 34 695 samples (16% of samples) have reported totals for garnets (eclogite and pyrope garnets). Individually pyrope and eclogite garnets are reported in ~13% and ~6% of samples respectively. These reporting percentages are consistent with those of other KIMs within KIDD (Table 2). Given the clear focus of KIMs for mineral exploration, it is unclear why grain counts are so sporadically entered. This may reflect incomplete integration of assessment report data within KIDD during one of the phases of data entry. Alternatively, and perhaps additionally, it may reflect selective data reporting by companies. For example, pyrope garnets are used as proxies for the diamondiferous potential of a kimberlite pipe and these data may be selectively reported. As well, low reporting of garnets could reflect the presence of pseudo-KIM's, that is, indicator grains derived from other crustal sources.

Where grain counts are reported and properly accounted for, the quality of data entry and integrity is typically high. For example, using a random subset of 24 assessment reports (totalling 8 683 samples), preliminary comparisons between total grain counts entered in KIDD and in the corresponding assessment report reveals a range of correspondence of ~87-100% (depending upon mineral species), However, a subset of the 24 assessment reports show a complete mismatch between the assessment report and KIDD (Figure 1).

Perfect correspondence between KIDD and an assessment report indicates that all reported indicator mineral totals are the same as the data source. Cases of 0% correspondence indicate either an absence of data in KIDD, errors in optical character recognition (OCR), or more frequently, significant errors in data entry which compromise summations of indicator minerals counts (Table 3).

4.1.2. Data limitations

The absence of sample weight is the greatest limitation of KIDD. This prevents any normalization of indicator mineral grain counts, thus limiting the ability to make cross-survey comparisons or join datasets from distinct surveys. Examination of assessment reports associated with archived samples in KIDD, reveals that sample weight/volume data are in fact reported, in varying ways, in the methodological descriptions of an assessment report. Reporting of these data can vary from precise measurements of sampled material, to approximations of sampled weights based on a known volume (e.g. 20L sample pail). Although often imprecise, these data can offer 'first order' (or better) approximations of the sampled weights and allow for rudimentary standardization of IM grain counts across surveys. If the volume is known, the mass can be approximated +/- 15%. Thus, limitations associated with sample weights/volumes lie not in the absence of information, but in the KIDD structure that does not contain a field for recording sample weight/volume data.

From the analysis of KIM data, and a sub-sample of assessment reports, it is apparent that some data quality issues exist within KIDD, due to the structure of the datasets. These issues are, however, not fatal for the usability of KIDD. When data in KIDD is complete it can be reliably used for analysis. In cases of data discrepancies, the primary cause is missing data rather than erroneous data. This is a recurring issue with garnet data in KIDD. These limitations are easily overcome by cursory examination of the KIDD data. Incomplete datasets can be rapidly assessed for completeness and eliminated or completed with the corresponding assessment report(s). It is therefore apparent that KIDD provides a useable data archive, though it requires data quality assessment prior to analysis.

Table 2: Summary of kimberlite indicator mineral fields within KIDD (total number of samples: 219 770).

Field	Description	Zero entries	Non-zero entries	Non-zero entries (%)	Null Values	Total Grains
DIAMOND	Count of identified diamonds in analyzed sample	219731	39	0.02	0	84
PYROPE_P	Count of identified pedridotite garnets in analyzed sample	191630	28140	12.80	0	311124
PYROPE_E	Count of identified eclogite garnets in analyzed sample	206046	13724	6.24	0	105685
TOT_GARN	Total garnet count (sum of PYROPE_P and	185075	34695	15.79	0	416809
CHRMDIOP	Count of identified chrome diopside in analyzed sample	193264	26506	12.06	0	258100
CHROM_SPIN	Count of identified chrome spinel in analyzed sample	200547	19223	8.75	0	221285
ILMN_PICRO	Count of identified ilmenite and picroilmenite (Mg-rich) in analyzed sample	196401	23369	10.63	0	596496
OPX	Count of identified orthopyroxene in analyzed	215166	4604	2.09	1	30395
OLIVINE	Count of identified olivine in analyzed sample	198814	20956	9.54	0	963382
TOTAL_GRAI	Sum of all grain count fields	145336	74434	33.87	1	2486551

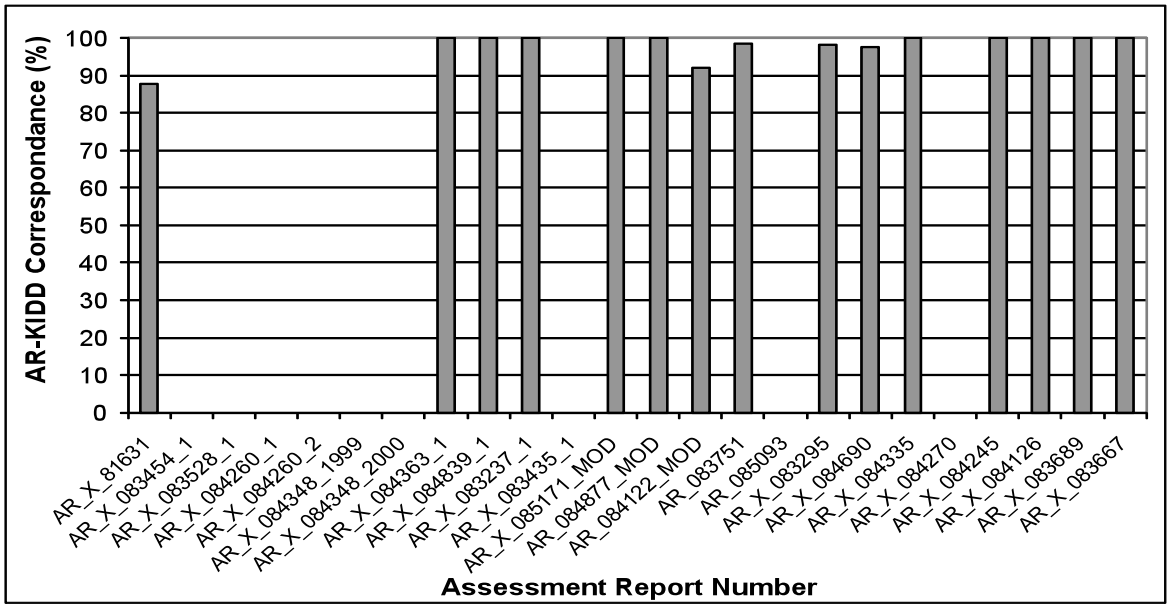

Figure 1: Histogram comparing the correspondence of data entered in KIDD vs. the data submitted as part of the corresponding assessment report (from a random subset of 24 assessment reports totalling 8683 samples).

5. Usability of KIDD: a graphical example

KIDD samples that are accurately reported allow for informative graphical plotting of datasets and, in some cases, identification of dispersal plumes. Reported samples allow for recognition of variable sampling density within the landscape (Figure 2). Higher density sampling is commonly used to infill broader 'property-scale' sampling efforts to constrain potential dispersal plumes. These variable scales of sampling can be identified from the KIDD data (Figure 2).

As well, reported KIM data are sufficient to recognize general patterns and spatial changes in KIM concentrations along a dispersal plume (Figure 3). However, the quality and spatial density of reported data within KIDD is variable. Not every reported survey allows for plotting detailed dispersal plumes. A persistent limitation in the use of KIDD within an archival data mining context is associated with unreported data. Using only the data contained within assessment reports, it is currently difficult to evaluate whether or not additional data were collected as part of sampling campaigns. In many cases, there is a likelihood that only a portion of all data are reported – generally sufficient to satisfy Canadian mining regulations. However, the percentage of unreported data is unknown and may, in some cases, limit the usability of KIDD.

Table 3: Correspondence statistics and error sources for cases of sub-100% correspondence between assessment reports and

Assessment Report Number	Correct entries/total	Incorrect entries	Correspondence percentage	Missing samples	Error sources
AR_X_81631	57/65	8	87.7	0	Incorrect entries, usually a 0 value, entered in KIDD or in assessment report
AR_X_084348_1999	0/304	304	0	1	Ilmenite counts entered as 0 in KIDD but have values in assessment report. No entries for total garnet counts in KIDD. One sample missing in KIDD
AR_X_084348_2000	0/304	304	0	76	Blank entries for ilmenite in KIDD. No entries for total garnet counts in KIDD. 76 samples from assessment report not entered in KIDD
AR_X_083435_1	0/145	144	0	1	Some correct entries for individual grain counts (chrome spinel, ilmenite-picroilmenite). Other indicator minerals not entered resulting is summation errors
AR_084122_MOD	2941/3194	253	92.08	0	Garnet totals missing affecting some total grain counts
AR_083751	225/228	3	98.68	0	Switching of two samples due to OCR error. One erroneous entry.
AR_085093	0/409	409	0	0	Two correct entries. Total counts in KIDD different from those reported in assessment report. Summation errors in original assessment report
AR_X_083295	54/55	1	98.18	0	Possible OCR error
AR_X_084690	42/43	0	97.67	1	Sample not entered in KIDD
AR_X_084270	0/157	157	0	0	Olivine grain counts not enetred in KIDD resulting in summation errors

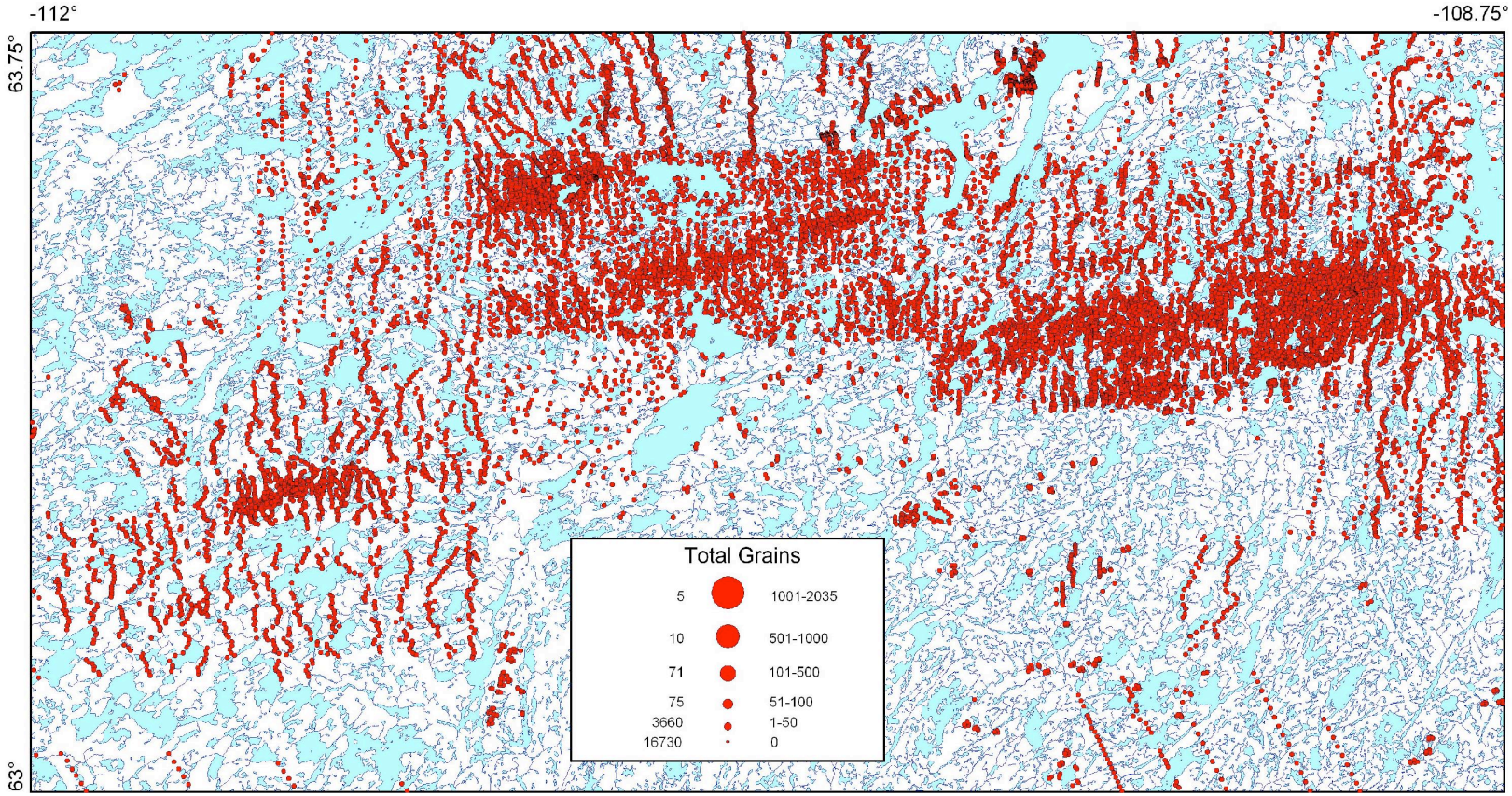

Figure 2: Example plot of KIDD data from the SE Slave kimberlite field, (e.g., Snap lake and Gacho Kue kimberlites). Within the map area, total grains are plotted to reveal variable spatial density of samples and clusters of sample points outlining potential dispersal plumes.

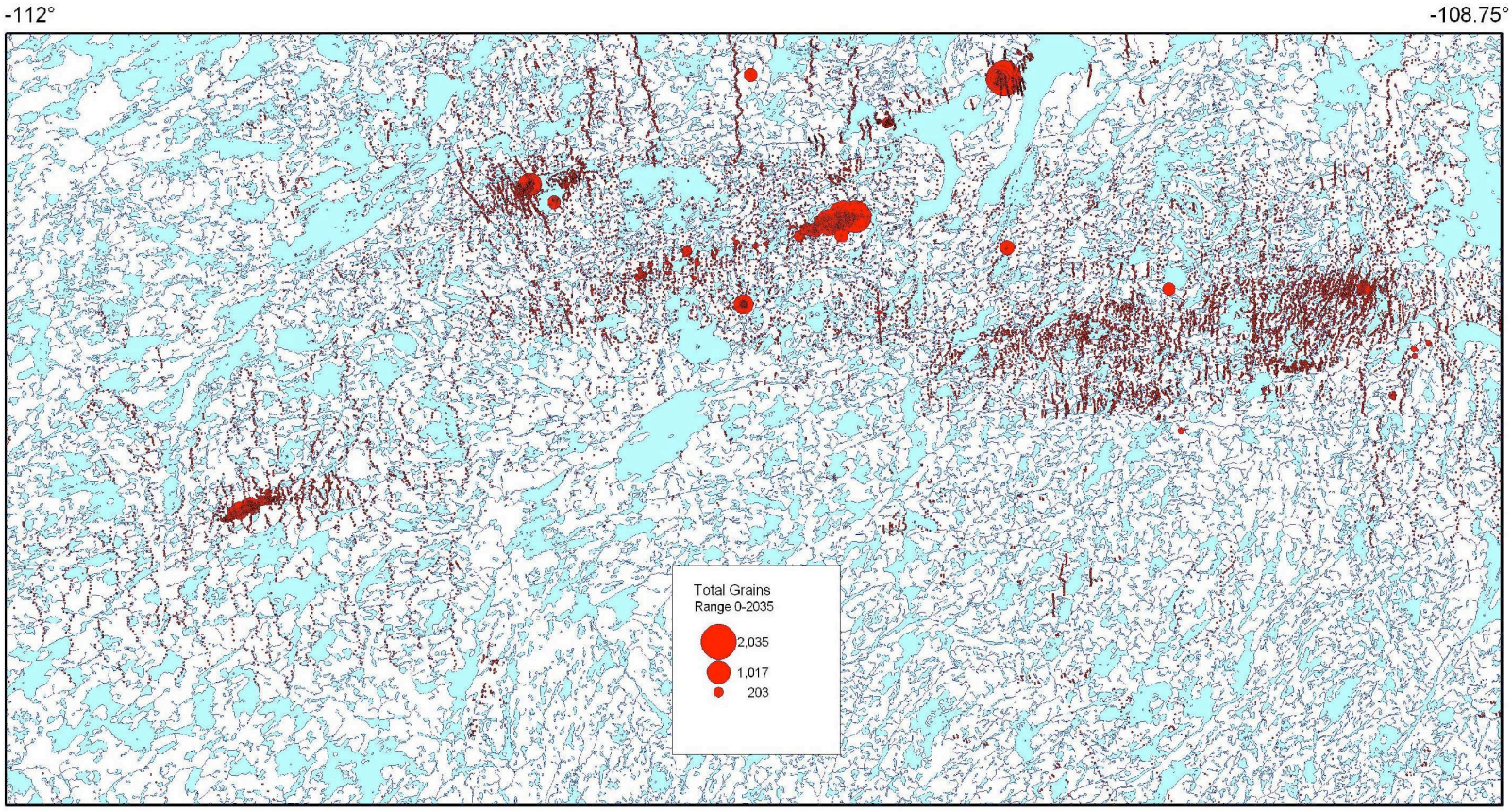

Figure 3: Example plot of KIDD data. Within the map area, total grains are plotted to reveal the variability of total grain counts outlining potential dispersal plumes.

5.1. Analytical Information: structure, content, and limitations

Analytical information within KIDD comprises three fields consisting of information on the upper and lower limits of the analyzed grains size fraction (SIZEFRACLO and SIZEFRACUP), and information on the analytical techniques used to treat samples (METHOD) (Table 4).

Table 4: Fields associated with grain size reporting and grain processing within KIDD

Field	Description
SIZEFRACLO	Lower limit(s) of analyzed grain size interval, units unspecified and variable.
SIZEFRACUP	Upper limit(s) of analyzed grain size interval, units unspecified and variable
METHOD	Analytical method(s) used in sample treatment

Variable reporting of the analyzed size fraction(s) is frequent and may reflect the variable focus of surveys, variable sample designs, and different analytical protocols. Nonetheless, quality of reporting of grain size data varies within KIDD. Over 67.5% of entries contain information on analyzed grain sizes (Table 5). Conversely, 30.9% of samples contain no grain size information and instead contain 0 or 9.99 entries that are assumed to be null values (though they do not respect the established protocol of Armstrong and Lee (2000;Table 5). Partial reporting of grain sizes (upper ranges only) accounts for 1.5 % of entries and include 0 (null value) as a lower grain size range. Lastly, 10.1% of entries reporting grain size information have inverted upper and lower ranges (Table 5).

KIDD does not specify grain size units. Grain sizes are assumed to be in mm based on the range and reporting format of values in KIDD. However, this creates a potential source of confusion if missing values such as -1 (Armstrong and Lee, 2000), and *de facto* missing values such as 0 and 9 (see section 3.4) are erroneously assumed to represent grain sizes on the Krumbein scale (phi scale - Krumbein and Sloss,1963), which correspond to some of the grain sizes commonly examined for indicator minerals

(very coarse sand, coarse sand; Wentworth, 1922), or for some geochemical analysis (clay).

Table 5: Summary of entries pertaining to grain size analysis within KIDD

Number of samples	Field entry	Percentage of entries
148 374	Complete grain size information	67.6
67 879	0 or 9.99 (assumed to be null values)	30.9
3 515	Only upper range of grain size range indicated, lower range commonly entered as 0 (null value)	1.5
15 049	Complete grain size information with inverted upper and lower ranges	10.1 *

* Calculated from total number of samples containing grain size information (148 374 samples)

Field entries pertaining to analytical methods contain a mixture of numeric and numeric-text entries. Over 32% of entries consist of either '0' or '9' values and contain no details on analytical methods (Table 5). These entries appear to be equivalent to null values though they do not conform to the protocol of Armstrong and Lee (2000) (see section 3.4 for further details). These null values also precede every text entry describing analytical methods (Table 4). Their recurrence suggests a systematic introduction in the database structure though it is unclear if these null value entries are artefacts of OCR treatment, conversion errors in the transition from individual spreadsheets to a database (see section 3.2), or a result of some as-yet unidentified error source.

In terms of methodological descriptions within the remaining 67.8% of entries, an absence of standardized reporting formats leads to 426 distinct methodological descriptions (occurring between one and 10,684 times). These entries comprise a varied mixture of details describing sample processing (e.g. crushing, 'desliming', milling), grain size processing (e.g. wet and dry sieving, sieved grain sizes), and heavy mineral separation techniques (e.g. magnetic separator, heavy liquid separation and specific gravity of separating fluid).

Table 6: Summary of field entries pertaining to analytical methods within KIDD

Number of samples	Field entry	Percentage of entries
5 342	0	2.4
65 310	9	29.8
2 568	9 followed by methods description	1.1
146 547	0 followed by methods description	66.7

5.2. Site descriptions: structure, content, limitations and improvements.

The COMMENTSAM field contains information about the sample, sample site characteristics, sample media, etc (J. Armstrong, Pers. Comm., 2013). Within KIDD, the COMMENTSAM field is used as a catch-all field incorporating a broad range of raw data, including analytical/sampling metadata, sampling site descriptions, and descriptions of the landscape context of sampling locations. The absence of standard reporting formats, and the mixed-purpose of the COMMENTSAM field result in 219,764 distinct entries. These include mainly information about sampling site characteristics or the characteristics of the sampled material. Significantly, the COMMENTSAM field also has information about analytical protocols that may or may not be duplicates of the entries under the METHOD field, information on the identity of the analytical lab, and notes and details on grain counts and noteworthy KIM grains. The absence of reporting protocols or formats complicates database queries and increases the chances that samples will be missed during queries based on site/sample characteristics.

5.2.1. Structure and content of site/sample descriptions

Site/sample descriptions are often the only source of data that shed light on the characteristics of the sampled media, or the landscape context of samples. These types of data may be critical to locating sample sites (given limitations with coordinates reporting, see section 4.1). They are also central to developing improved understanding of glacial processes, and to detailed understanding of sampling site characteristics in the context of surficial geology maps. Therefore, we use the field entries associated with the word ‘till’ to examine the variability of site and sample descriptions in the COMMENTSAM field.

Four reporting formats seem to dominate descriptions of site characteristics within the COMMENTSAM field. These formats reveal increasing levels of details about sample/site characteristics. They include:

1) *Single descriptors*: the sampled material is described by a single word or technical term. Although they offer the least amount of information on sample/site characteristics, these are the dominant types of entries in KIDD and account for 29,977 entries (Table 7).

2) *Compound descriptors*: a multi-word entry describes the sampled material (e.g. undifferentiated till, Till/Diamicton). Compound descriptors can lead to more detailed site characterization than single descriptors as they refer to the sedimentology of the sampled media (8,260 entries), or the landform context of samples (2,071 entries). Examples of this structure include: 'Till Veneer', 'Till: Inactive Frost Boil' (Table 7).

3) *Single descriptors and qualifier(s)*: a single word entry (273 entries) describes the material and is followed by one or more qualifiers describing the sample sedimentology (118 entries), the landform context (65 entries), the material state (73 entries), or some other qualifier (1 entry) (Table 7). Examples of this structure include: 'Till, ablation', 'Till, on coastal plain', or 'Till pocket; in boulder field' (Table 7).
Although the *Compound descriptor* structure and the *Single descriptor and qualifier* structures can seem similar, they are differentiated by the structure of the entry, and the fact that *Single descriptor and qualifier* structures are often more detailed and informative.

4) *Compound descriptor and qualifier(s):* a multi-word entry describes the material and qualifies the characteristics by describing the sedimentology, landscape context, and/or the character of the material (6 entries). An example of this structure can be: 'till veneer; by outcrop; large boulders around; sandy till' (Table 7).

5) Erroneous entries due to typos or spelling errors are relatively rare in the COMMENTSAM field as they account for only 24 entries out of the 219 764 entries incorporating the word till. Individual companies (or samplers) have distinctive reporting styles (it is unclear if these styles follow any protocol so we describe them as styles rather than formats). Therefore, many entries have a similar style and the frequency of entries in Table 7 may give a false impression of the diversity of entries. This realisation

is important in regards to attempts at standardizing the COMMENTSAM field: the apparent complexity of this field lies in its irregular reporting structure, not necessarily in the complexity of the materials encountered during sampling. Third, we have illustrated the structure of this field by using the word 'till' as the lead field entry. A similar process could be performed with the word 'esker' as a lead entry. However, given the clear focus on till sampling in diamond exploration, we consider that the chosen example adequately captures and illustrates the diversity and structure of this field.

5.2.2. Limitations and improvements to the COMMENTSAM field

Clarifying and standardizing the structure and content of the COMMENTSAM field could potentially yield some of the greatest improvements in the usability of KIDD. Site and sample data are highly valuable and informative data that bear heavily on data and process interpretations when trying to interpret archival data within a landscape context of IM dispersal. Some potential improvements in structure and content are listed below:

1) *Distinguishing sample and site characteristics:* In its current form, the COMMENTSAM field may include both sample and site characteristics. These two descriptors should not be viewed as interchangeable entries but as distinct and complimentary entries.

2) *The need for a hierarchical descriptive approach*: To facilitate integration into a database structure, both sample and site descriptions need to be hierarchical with clear identification of a primary descriptor, followed by secondary and tertiary descriptors. Each descriptor could also include some kind of modifier, as long as it occupies a single database field. This structure does not preclude an open field to allow for unique entries that do not fit pre-existing field descriptors. This option, however, should be available after main materials/site descriptors have characterized the site.

Table 7: Variability and types of reporting formats of the term 'till' within the COMMENTSAM field of KIDD

Entry types	Materials [n]	Compound Descriptors [n] Sedimentology	Landform	Material state	Other
Single descriptor	Till [20557]				
	till [6153]				
	TILL [3157]				
	Till? [68]				
	till? [39]				
	TILL? [2]				
	TILLSHOR [1]				
Compound descriptors		Undifferentiated Till [5088]	Till/Frost Boil [1094]	TILL-DRY [24]	Till sample [42]
		undifferentiated till [1897]	Till Boil [376]	till slush [1]	Till/Lee Side [6]
		Till/Diamicton [1176]	Till Veneer [227]		Till trapped between low outcrop [3]
		Undifferentiated till [46]	Till: inactive frostboil [139]		
		Till Material [39]	Till: active frostboil [45]		till slope [1]
		TILL/GF [6]	till veneer [34]		Till/outcrop [1]
		TILL-GF [3]	Till: Inactive Frost Boil [33]		Till; High area [1]
		Till or Glaciofluvial [2]	Till Frost Boil [19]		
		Till or glaciofluvial [1]	Till-Shoreline [18]		
		till outwash [1]	Till shoreline [15]		
		TILL/Glacio-Fluvial [1]	Till/Inactive Frost Boil [15]		
			Till/Beach [12]		
			Till or Esker [7]		
			Till plain [7]		
			till/esker [5]		
			Till mound [4]		
			Till/Esker [3]		
			Till/Esker? [3]		
			till/lacustrine [3]		
			till/beach? [2]		
			till plain [2]		
			till; found rock [2]		
			Till: Active Frost Boil [1]		
			till; esker [1]		
			TILL-ESKER? [1]		
			Till: sediment blanket [1]		
			till blanket [1]		
			till moraine [1]		

Table 7 (continued): Variability and types of reporting formats of the term 'till' within the COMMENTSAM field of KIDD

Entry Types	Material [n]	Qualifiers [n] Sedimentology	Landform	Material State	Other
Single Descriptor + Qualifier(s)	Till [184]	and frost boils [62]	Shoreline [18]	(dry) [54]	N sloping hill [1]
	TILL [82]	outwash [7]	Beach [12]	(Dry) [9]	
	till [3]	(w lacustrine on top?) [6]	local Esker [5]	(wet) [5]	
	TILL? [3]	shells [6]	low ground; raised beaches [4]	DRY [3]	
	Till (?) [1]	Ablation [4]	rolling upland [3]	poor sample [1]	
		moraine? [4]	beach [2]	sample quality below average [1]	
		GF [4]	beach? [2]		
		outwash - unsorted [4]	beach (very sandy) [2]		
		Possibly GF [3]	on plains [2]		
		lacustrine [3]	ESKER? [1]		
		lacustrine? [3]	Esker like [1]		
		Glaciofluvial? [2]	SHOR [1]		
		Soil? [2]	on coastal plain [1]		
		Glacio-Fluvial? [1]	fluvial component [1]		
		Sandy [1]	reworked fluvial [1]		
		STONE [1]	raised beaches [1]		
		outwash fines [1]	side of plateau [1]		
		clay plus limestone [1]	top of ridge [1]		
		sandy/silt [1]	low ground [1]		
		stream sediment [1]	low ground: rocky outcrops [1]		
		GF? [1]	till; edge of hills; raised beaches [1]		
			till pocket; frost boil [1]		
			till pocket; in boulder field [1]		
			till pocket with rock [1]		
Compound descriptor + qualifier(s)	till veneer; by outcrop; large boulders around; sandy till [1]			Entry errors (typos, OCR)	Tilt [6]
	till veneer; near outcrop; W of lake; sandy till; surrounded by boulders [1]				tiil [3]
	till veneer; on outcrop; N of lake; sandy sample [1]				TIll [2]
	till veneer; sandy (washed?); surrounded by large boulders [1]				TIL BOIL [2]
	till veneer; washed out - sandy/silty; outcrop ridge to N; surrounded by boulders [1]				Til! [2]
	till with clay on bedrock [1]				Till [2]
					tilt [2]
					til! [1]
					til] [1]
					tii! [1]
					TILL/B [1]
					TILL/BOI [1]

6. Summary of known KIDD limitation and recommendations for maximizing the potential of KIDD

From examination of the structure and content of KIDD, key limitations have been identified. In most cases, limitations are related to the structure of KIDD and therefore require fairly simple solutions. Key limitations and improvements for maximizing the potential of KIDD are listed below:

1) *Standardizing and resolving coordinate reporting formats*: Despite the fact that Armstrong and Lee (2000) established reporting and conversion protocols for reporting geographic coordinates within KIDD, many assessment reports do not specify the reference datum within assessment reports. In some cases, displayed and reported data use a different geographic datum. In other cases, these details are not provided and it is unclear how the data are reported. Clarifying these basic data reporting requirements is crucial to properly locating sampling site, especially in an archival data context where original field notes are not available.

Recommendation: Add additional fields for Datum and conversion information along with a protocol for identifying an absence of reported information.

2) Null value clarification: Armstrong and Lee (2000) established a null value reporting protocol. This protocol is inconsistently used within KIDD and, in places, *de facto* null values are used but do not conform to the original protocol. As well, the existing protocol of Armstrong and Lee (2000) (-1 values) is potentially confusing when used in grain size fields.

Recommendation: Initiation of a more universally accepted and recognized null value entry (e.g. -9999) for KIDD to avoid potential confusion and to clearly differentiate null value entries from others.

3) Sample weight reporting: Reporting of sample weights is critical for normalization of reported KIM grain counts. It is also essential to any cross-survey comparisons. The lack of sample weight reporting stems mainly from the KIDD structure, which does not include a field for reporting these data, despite the fact that many assessment reports do report sample weights, to varying degrees of precision. Sample weights need to be reported in a database such as KIDD to maximize the

utility of data analysis in an archival context. At a minimum, this requires addition of a sample weight field. Even rudimentary reporting of sample weight measurements can be used to normalize grain count data, understanding that a degree of uncertainty will be associated with the normalization.

Recommendation: An addition to KIDD could include a sample weight field, and one or two additional fields used to describe the way in which the data were reported in the assessment report. For example, if available, a direct weight could be reported in a first field (e.g. 20 kg). In cases where the weight is either not directly reported or is estimated, secondary and tertiary fields could provide information on accuracy of the measurement (e.g. weight measured vs. estimated), and where applicable, a field describing the estimation approach (e.g. weight estimated from volume of sampling pail).

4) *Improved sample and site descriptions:* The use of KIDD as an archival tool requires robust and informative descriptions of sampled material and sample site characteristics. Again, insufficient fields within the KIDD structure preclude adequate characterization of sampled material and sample sites. This limits the usability of KIDD. Although many of these sample/site descriptions are reported in KIDD, the inadequate structure strongly limits the potential of KIDD as an archival tool.

Recommendation: A more hierarchical descriptive structure would help maximize the potential of KIDD.

Beyond the original intention of KIDD as an archive for KIM data, revising and improving the existing KIDD structure could serve as a leading example and as a useable template for development of parallel databases (or expanded versions of KIDD) as new commodities emerge as viable exploration targets.

6.1. Tools developed for the assessment of KIDD

Much of the KIDD analysis and descriptive statistics within this report were extracted using a series of computer scripts. These scripts have been collated into three useable tools that run as part of ArcGIS and that can be integrated as a ‘Toolbox’ within ArcGIS software suite. They are scripted in the Python (version 2.7) programming language. These tools are available to interested parties through a request to the senior author and

allow for further analysis of the integrity of KIDD. Their basic structure and operation are briefly described below.

1) ***KIDDZero_orNot.py*** is a script used for a first order assessment of KIDD and is capable of extracting summary statistics of the KIDD entries and to identify zero/non-zero entries within grain count fields. It was used to generate Table 2. This tool also identifies potential erroneous and/or missing value entries (e.g. '-1' values; see section 3.4). The operational sequence consists of a sequential assessment of KIDD fields and identification of single '0' entries. Summary statistics are calculated from identification of '0' entries (Table 2).

2) ***GrainCountTotalsCheck.py*** *is a script used to assess grain count totals (in particular for the TOT_GARN field which sums both pyrope and eclogite garnet, and for TOTAL_GRAIN field which sums all grain count fields. The operational sequence is as follows:*

   i. *A cursor identifies and sums all values within a row.*
   ii. *The calculated sum is compared to the reported sum (i.e. TOT_GARN or TOT_GRAIN)* in KIDD and identifies matching/non-matching cells.
   iii. *Erroneous entries (if any) are flagged*

   Use of this tool highlighted the fact that all calculated sums matched with those reported in KIDD. This complete match probably indicates that reported sums in KIDD were calculated automatically from the grain count entries, rather than entered manually from assessment reports. Thus, this tool cannot assess the integrity of the data entry. Any missing values or data entry errors (e.g. '-1' values) are included within the calculations.

3) ***ArKIDDCompare.py*** is a script designed to assess the integrity of data entry by comparing entries in KIDD to the data originally submitted in the assessment reports. Implementation requires extraction and OCR treatment of the assessment report data into an MSExcel spreadsheet. The spreadsheet is

imported into ArcGIS where it can be compared to the KIDD. The operational sequence is as follows:

i) The tool matches sample numbers between KIDD and an assessment report.

ii) The data entered in each matched row are compared and matching/non-matching entries in each grain count field are identified.

iii) An output table (Fig. 4) is generated and shows which samples have matching grain counts, and which do not. If a sample number is missing from KIDD, then a 'missing sample' entry occurs in the output table.

Common identified errors include OCR errors where the original scan of the assessment report is of low quality. In other cases, manual data entry errors occur in KIDD, possibly resulting from assessment reports where OCR could not be used. Overall, this tool allows for rapid identification of errors within KIDD and the location of these errors within the database. This tool was used extensively to generate the KIDD integrity statistics presented in section 4.2.

AR_081631_K_CD

OID	SAMPLE_NO	SAMPLE_NUM	CD	CHRMDIOP	MATCHED
7	B13	B13	0	1	no match
14	B21	B21	0	1	no match
52	B73	B73	1	0	no match
0	B6	B6	1	1	match
1	B7	B7	0	0	match
2	B8	B8	6	6	match
3	B9	B9	2	2	match
4	B10	B10	0	0	match
5	B11	B11	0	0	match
6	B12	B12	1	1	match
8	B14	B14	1	1	match

Figure 4: Portion of an output table generated using the ArKIDDCompare.py tool. The grain count fields of matched sample numbers from KIDD and a spreadsheet of the assessment report data are compared and flagged for matching/non-matching entries.

7. Conclusions

- The Kimberlite Indicator and Diamond Database (KIDD) is a relational database archiving grain count data from Kimberlite Indicator minerals.

- A number of limitations of KIDD have been identified. Some create minor limitations to the use of KIDD, while others potentially create more serious limitations. Importantly, none are insurmountable obstacles to the use of KIDD in an archival context.

- Minor limitations to the use of KIDD are associated with non-adherence to protocols for reporting null values within KIDD. These limitations can generally be identified with detailed examination of the KIDD data and associated assessment reports. However, clarification and adherence to the null value reporting protocol would simplify end-user access to KIDD.

- More serious limitations to using KIDD stem from an absence of clear reporting of geographic datum of sample site coordinates. These have potential repercussions when trying to locate existing sample sites archived in KIDD.

- Limitations also occur as a result of irregular and non-standardized reporting of sample weights within KIDD. This limits normalization of data, and cross-survey comparisons. This limitation stems from an absence of a dedicated field within the KIDD structure rather than from an absence of these data within assessment reports.

- Non-standardized and conflation of sampled media and sample site descriptions limits the usability of KIDD in an archival context. Some of these descriptive data are available. Inadequate reporting protocols and structures (i.e. dedicated fields) within KIDD currently limit the ability of end-users to query KIDD for information on sample/site characteristics.

- Limitations in the structure of KIDD do not fundamentally undermine the quality of data within KIDD. These data are fully usable as exemplified by assessment of a subset of 24 assessment reports (8683 samples), which importantly show that most reported data within KIDD faithfully replicate the original reported data in assessment reports.

- Introduced errors associated with data entry during development of KIDD seem to be few and far between and do not compromise the usability of KIDD.

- KIDD is therefore a reliable and useable dataset, though end-users need to understand its limitations.

- Improvements to the KIDD structure could yield large gains for KIDD users and would greatly decrease the need of end-users to refer to assessment reports in order to qualitatively and quantitatively use KIDD data. This would support improvements to mineral exploration infrastructure, which still lags behind numerous other jurisdictions in Canada (e.g. LookNorth 2012, p. 14).

- Data capture to KIDD and data quality assessment and correction suffer from the age old problem of adequate funding that were an issue in the 1980's (e.g. Simpson, 1985) and continue to be an issue in the 21st century (Duke, 2010).

8. Acknowledgments

Discussion of the origin of KIDD with John Armstrong was very beneficial. Input by NWTGeoscience office staff J. Ketchum, and B Elliot are much appreciated. Discussion and comments on a draft document by R.D. Knight, D.R. Sharpe helped to focus the work flow progression. Internal review by S. Adcock and E. Grunsky helped clarify the text. This work is funded as part of the Targeted Geosciences Initiative IV (TGI-IV) Program: Enhanced Effectiveness of Deep Exploration, and specifically the Methodological Project on Indicator Mineral Dispersal.

9. References

Armstrong, J.P. 2003. Diamond Discovery in the Slave Craton: Compilations of Exploration Data as Tools for Future Discovery. 8th International Kimberlite Conference, June 22-27, 2003, Victoria, BC., Long Abstract, 5 p.

Armstrong, J.P. and Lee, C.A. 2000. Kimberlite Indicator and Diamond Database (KIDD): A compilation of publically available till sample locations and kimberlite indicator mineral picking results, Slave Craton and environs, Northwest Territories and Nunavut, Canada. DIAND NWT Geology Division, EGS2000-03.

Armstrong J.P., Skinner S., Mepham J., Cairns S.R. 2004. Kimberlite Indicator and Diamond Database (KIDD) Update: A compilation of publically available till sample locations and kimberlite indicator mineral picking results, Slave Craton and environs, Northwest Territories and Nunavut, Canada. NWT Open Report 2004-004.

Australia Victoria Mineral Exploration Geochemistry Data, Accessed 2013-07-31; ttp://www.dpi.vic.gov.au/earth-resources/about-earth-resources/projects/mineral-exploration-geochemistry-database

Krumbein, W.C., Sloss, L.L. 1963. Stratigraphy and sedimentation, San Francisco, W. H. Freeman.

Duke, J.M. 2010. Government geoscience to support mineral exploration: public policy rationale and impact; prepared for Prospectors and Developers Association of Canada, 72 p.

Jones, K. 20 . A New Digital Diamond Indicator Mineral Database for the Slave Craton NWT; downloaded 2013-09-05; http://www.geosoft.com/media/uploads/resources/success-stories/gde_cs_2008_01_web.pdf

LookNorth 2012. Mining Sector Assessment Mining in the Canadian North: opportunities for remote sensing technologies to add value to the sector; dec 2012; downloaded 2013-07-31; http://www.looknorth.org/wp-content/uploads/2011/07/LOOKNorth-Mining-Sector-Assessment-Dec-2012-compressed.pdf

Keller, G.R. and Bogdan, D.J. 2004: The Manitoba Kimberlite Indicator Mineral Database: an update; in Report of Activities 2004, Manitoba Industry, Economic Development and Mines, Manitoba Geological Survey, p. 320–322.

Ministry of Environment, Ontario. Accessed 2013 july 31. http://www.ene.gov.on.ca/environment/en/subject/wells/STDPROD_075977.html

Northwest Territories and Nunavut Mining Regulations 2013. C.R.C., c. 1516; Current to July 10, 2013; Last amended on July 28, 2008; Published by the Minister of Justice at the following address: http://laws-lois.justice.gc.ca

Northwest Territories Geoscience Office website (2013): http://ntgomap.nwtgeoscience.ca/.

NSW Industry and Investment, 2010. Exploration Reporting: A guide for reporting on exploration and prospecting in New South Wales, 32 p; http://www.industry.nsw.gov.au.

Paulen, R. 2012. Tri-territorial database : Indicator Minerals; 2012 Nunavut Mining Symposium – April 16-19, 2012; http://www.nrcan.gc.ca/earth-sciences/about/current-program/geomapping/minerals/7770, 2013-07-31.

PPDM Association, accessed 2013-07-31; http://www.ppdm.org/

Russell, H. A. J., T.A. Brennand, C. Logan, and D.R. Sharpe, 1998, Standardization and assessment of geological descriptions from water well records: Greater Toronto and Oak Ridges Moraine areas, southern Ontario: Current Research 1998-E: Geological Survey of Canada, 89-102.

Simpson, F. 1985. A Users' Guide to Core Storage Facilities in Canada, Geological Survey of Canada Paper 84-23.

Sweden Exploration reports, 2013. Accessed 2013-07-31; ttp://www.sgu.se/sgu/eng/samhalle/malm-mineral/mininfo/prosprapp_e.html