Elisabeth Mödden, Christa Schöning-Walter, Sandro Uhlmann

# Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek

Breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse

**Digitale Medienwerke machen inzwischen den größten Anteil des jährlichen Zugangs der Deutschen Nationalbibliothek (DNB) aus, mit steigender Tendenz. So sind die Bestände im Jahr 2016 um etwa 500 000 physische Medienwerke und 1,3 Millionen digitale Publikationen gewachsen, davon etwa 280 000 E-Books, Online-Hochschulschriften und Musikalien sowie mehr als eine Million elektronische Zeitschriftenartikel, E-Paper-Ausgaben und Zeitschnitte von Webseiten. Die insgesamt steigenden Mengen sind eine Herausforderung für das Sammeln und Verzeichnen. Zugleich bieten die Veränderungen neue Chancen für die Benutzung, indem beispielsweise auch einzelne Artikel gesucht und gefunden werden können.**

**Inhaltserschließung ermöglicht es, die großen Bestände für das Retrieval thematisch zu strukturieren. Die DNB beschäftigt sich seit einigen Jahren mit der Frage, wie sich die Prozesse der inhaltlichen Erschließung maschinell unterstützen lassen, um trotz neuer Medienformen und immer größerer Mengen zu erschließender Medieneinheiten eine möglichst einheitliche und vollständige Anreicherung mit inhaltsbeschreibenden Metadaten zu erreichen. Auch weitere Vorteile maschineller Prozesse, beispielsweise die Möglichkeit, bislang nicht berücksichtigte Gliederungsebenen wie die schon genannten Zeitschriftenartikel klassifikatorisch und verbal erschließen zu können, sollen konsequent genutzt werden.**

**Seit 2010 werden die digitalen Medienwerke in der DNB nicht mehr intellektuell, sondern zunehmend mit maschinellen Verfahren erschlossen.[1] Im September 2017 wurde die Anwendung maschineller Erschließungsverfahren erstmalig auf physische Medien ausgeweitet.[2,3] Im Strategischen Kompass 2025[4] der DNB und in den Strategischen Prioritäten[5] ist die Neuausrichtung der inhaltlichen Erschließung auch für die nächsten Jahre als ein wichtiges Handlungsfeld dargestellt. Dieser Beitrag beschreibt den Stand der Umsetzung und die weiteren Aufgaben.**

### Erschließungsmethoden

Die inhaltliche Erschließung in der DNB richtet sich nach der Zuordnung der Publikationen zu den Reihen der Deutschen Nationalbibliografie. Seit dem Bibliografiejahrgang 2004 erhält jede Publikation eine Sachgruppe, die thematische Gliederung in etwa hundert Sachgruppen folgt der Dewey-Dezimalklassifikation (DDC)[6]. Für die in der Reihe A verzeichneten Verlagspublikationen wird intellektuell auch eine Tiefenerschließung mit vollständigen Notationen der DDC sowie eine verbale Erschließung mit dem Schlagwortvokabular der Gemeinsamen Normdatei (GND)[7] durchgeführt.

Die Entwicklung maschineller Prozesse für die klassifikatorische und verbale Erschließung wurde im PETRUS-Projekt[8] begonnen. 2012 konnte die maschinelle Sachgruppenvergabe in Betrieb genommen werden, 2014 dann die maschinelle Schlagwortvergabe. Für medizinische Publikationen wurden 2015 erstmals verkürzte Notationen der DDC maschinell vergeben. Das Schema mit 140 medizinischen Kurznotationen war bereits Ende 2005 für die Erschließung medizinischer Dissertationen eingeführt worden. Zurzeit wird an einem Klassifikationsschema mit verkürzten DDC-Notationen für alle Fächer gearbeitet. Eine Systematik mit 72 Klassen für die Informatik wird gerade erprobt.

Im produktiven Erschließungsprozess werden die inhaltlich wesentlichen Begriffe einer Publikation mithilfe einer mehrstufigen linguistischen Analyse ermittelt und mit dem Schlagwortvokabular abgeglichen.

Für die maschinelle Klassifikation mit Sachgruppen und Kurznotationen verwendet die DNB ein maschinelles Lernverfahren.[9] Die sprachlichen Merkmale ausgewählter Textausschnitte und vorhandener Metadaten werden mit linguistischen und statistischen Methoden analysiert. In der Trainingsphase erstellt das System anhand intellektuell erschlossener Publikationen ein Referenzmodell für alle Klassen. Für die Modellbildung ist es wichtig, dass in jeder Klasse möglichst charakteristische Lernbeispiele in ausreichender Anzahl vorhanden sind. Im produktiven Erschließungsprozess errechnet das System dann ein statistisches Maß dafür, wie stark die Inhalte einer neuen Publikation mit den erlernten Mustern übereinstimmen. Die am besten passenden Sachgruppen und Kurznotationen werden der Publikation zur thematischen Einordnung als Metadaten zugeordnet.

Die maschinelle Schlagwortvergabe hingegen basiert ausschließlich auf linguistischen Verfahren.[10] Für die Analyse deutschsprachiger Texte sind etwa eine Million Terme aus der GND – Sachbegriffe, Personen, Geografika, Körperschaften, Kongresse und Werke – zusammen mit den vorhandenen semantischen Informationen als Schlagwortvokabular in die Erschließungssoftware integriert worden. Im produktiven Erschließungsprozess werden die inhaltlich wesentlichen Begriffe einer Publikation mithilfe einer mehrstufigen linguistischen

Analyse ermittelt und mit dem Schlagwortvokabular abgeglichen. Dabei müssen auch die vielen mehrdeutigen Begriffe der deutschen Sprache in den richtigen Bedeutungszusammenhang eingeordnet werden. Bei gleichlautenden Begriffen mit verschiedenen Bedeutungen wie beispielsweise »Bank« oder »Pfund« muss der Bezug zum richtigen Term im GND-Vokabular gefunden werden. Als Analyseergebnis werden schließlich bis zu sieben Schlagwörter pro Publikation ausgewählt, die entsprechenden Verknüpfungen mit den Datensätzen in der GND werden im Titeldatensatz verzeichnet. Damit kann dann auch die Normdatei mit ihren weitverzweigten Vernetzungen als Sucheinstieg für die maschinell erschlossenen Publikationen genutzt werden.

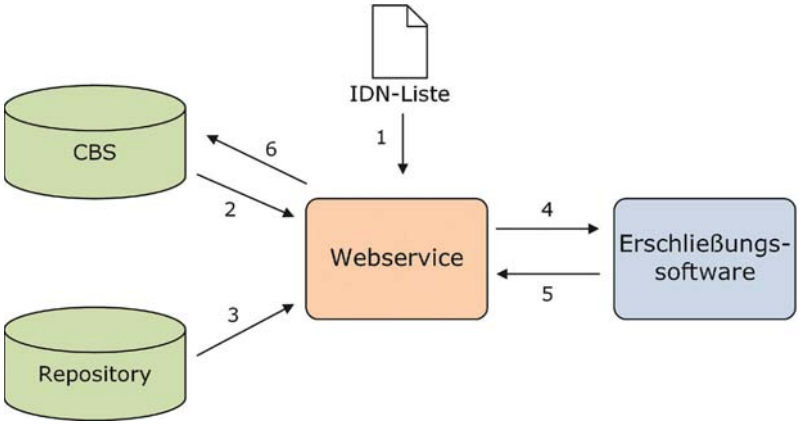


Abbildung 1: Technischer Ablauf der maschinellen Erschließung im Produktionsbetrieb der DNB.

Die Erschließungssoftware wurde in Zusammenarbeit mit dem Freiburger Unternehmen Averbis erstellt und ist in die Systeminfrastruktur der DNB eingebunden. Die maschinelle Klassifikation ist für die Sprachen Deutsch und Englisch implementiert, die Schlagwortvergabe ist bisher noch auf deutschsprachige Publikationen beschränkt.

### Prozessablauf

Im produktiven Betrieb startet die maschinelle Erschließung (siehe Abbildung 1) täglich automatisch zu einer festgelegten Zeit damit, dass eine Liste der neu zu verarbeitenden Publikationen [1] an einen Webservice übergeben wird. Dieser holt die schon vorhandenen Metadaten [2] aus der Katalogisierungsdatenbank (CBS) und die digitalen Volltexte oder Inhaltsverzeichnisse [3] aus dem Repository. Vor Übergabe an die Erschließungssoftware [4] werden die Speicherformate in einfache Textdateien umgewandelt und die vorwiegende Sprache der Publikation wird bestimmt. Die zurückgelieferten Analyseergebnisse [5] werden im Titeldatensatz der Publikation verzeichnet [6]. Auffälligkeiten im Verarbeitungsprozess werden in Systemdateien protokolliert.

Das Erschließungssystem bietet die Möglichkeit, verschiedene Konfigurationen einzurichten, damit Publikationen unterschiedlicher Art differenziert verarbeitet werden können. Dabei handelt es sich um Parametereinstellungen, die in Testreihen optimiert wurden. Bei der Sachgruppenvergabe wird auf diese Weise beispielsweise das Klassifikationsmodell definiert. Abhängig von den Publikationsmerkmalen wird im produktiven Betrieb eine bestimmte Konfiguration angesteuert: So werden digitale Monografien anders prozessiert als Zeitschriftenartikel, deutschsprachige Texte anders als englischsprachige, Volltexte anders als digitalisierte Inhaltsverzeichnisse.

Pflege und Weiterentwicklung der Software, der Trainingskorpora und des Schlagwortvokabulars führen stetig zu Verbesserungen des Gesamtsystems. Die intellektuell erschlossenen Medienwerke fließen zu bestimmten Zeitpunkten als neue Trainingsbeispiele in die Lernprozesse der Klassifikation mit ein. Auch das Vokabular für die maschinelle Schlagwortvergabe wird systematisch bearbeitet und künftig regelmäßig mit dem aktuellen Stand der GND abgeglichen. Bei maßgeblichen Fortschritten stellt sich jeweils die Frage, ob Erschließungsvorgänge wiederholt werden sollten. So ist die maschinelle Schlagwortvergabe bisher jährlich neu durchgeführt worden, nachdem

1 Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 20 - 22

2 Junger, Ulrike; Schwens, Ute: Die inhaltliche Erschließung des schriftlichen kulturellen Erbes auf dem Weg in die Zukunft. In: Dialog mit Bibliotheken, 29 (2017) 2, S. 4 - 7

3 http://www.dnb.de/inhaltserschliessung

4 Deutsche Nationalbibliothek 2025: Strategischer Kompass - Leipzig; Frankfurt, M.: Dt. Nationalbibliothek, 2016. Online unter https://d-nb.info/1112299254/34

5 Strategische Prioritäten 2017 – 2020 - Leipzig; Frankfurt, M.: Dt. Nationalbibliothek, 2016. Online unter https://d-nb.info/1126594776/34

6 http://www.dnb.de/Subsites/ddcdeutsch/DE/Home/home_node.html

7 http://www.dnb.de/gnd

8 Schöning-Walter, Christa: PETRUS – Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 15 - 19

9 Mödden, Elisabeth; Tomanek, Katrin: Maschinelle Sachgruppenvergabe für Netzpublikationen. In: Dialog mit Bibliotheken, 24 (2012) 1, S. 17 - 24

10 Uhlmann, Sandro: Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei. In: Dialog mit Bibliotheken, 25 (2013) 2, S. 26 - 36

11 http://id.loc.gov/authorities/subjects

12 http://wiki.dbpedia.org/

13 https://www.mpi-inf.mpg.de/yago

das Vokabular optimiert wurde. Die Möglichkeit der zyklischen Wiederholung soll künftig systematisch dafür genutzt werden, die Qualität der maschinell vergebenen Metadaten zu verbessern und rückwirkend auch Publikationsgruppen mit zu erschließen, die bisher noch nicht berücksichtigt werden konnten.

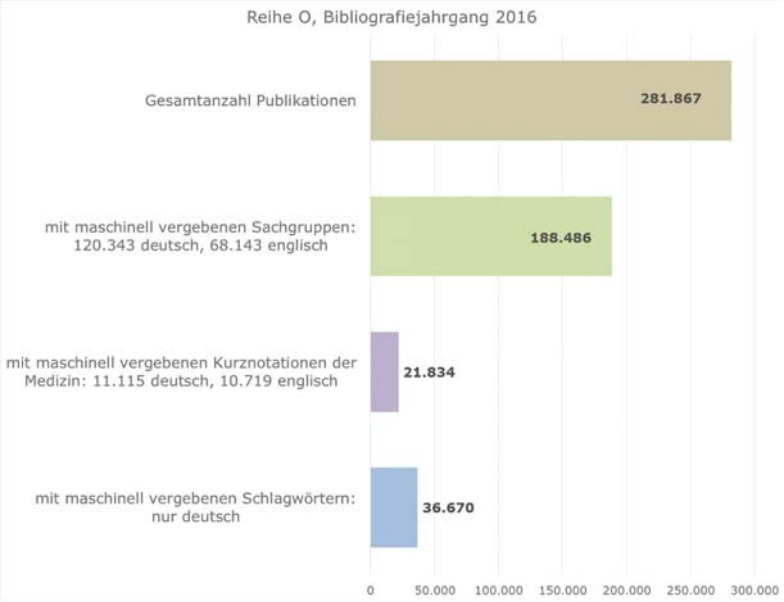


Abbildung 2: Anzahl der monografischen Netzpublikationen mit maschineller Erschließung im Verhältnis zur Gesamtzahl der 2016 in der Reihe O verzeichneten Monografien.

### Anwendungsbereiche

Anfang 2010 wurde die Reihe O zur nationalbibliografischen Verzeichnung der digitalen Medienwerke, auch als Netzpublikationen bezeichnet, eingeführt. Der Einsatz der maschinellen Erschließungsverfahren war zunächst ausschließlich auf die digitalen Monografien ausgerichtet. Etwa 65 Prozent dieser Titel erhalten mittlerweile eine Sachgruppe durch maschinelle Textanalyse. Etwa 35 Prozent werden nicht prozessiert, weil es sich um Titel der Belletristik handelt oder um Publikationen in anderen Sprachen als Deutsch oder Englisch. Für die Belletristik liefern die Analyseverfahren bisher noch keine sinnvollen Ergebnisse. Falls keine eigenen Metadaten erzeugt werden können, nutzt die DNB die mitgelieferten Fremddaten.

Abbildung 2 zeigt für 2016 den Anteil der Monografien in der Reihe O, der maschinell erschlossen wurde. Die Vergabe von Kurznotationen beschränkt sich zurzeit noch auf die Medizin. Für die Schlagwortvergabe sind erste Konfigurationen implementiert, und zwar für deutschsprachige Hochschulschriften, Publikationen verschiedener Wissenschafts- und Universitätsverlage, Aufsätze aus dem akademischen Bereich und für die inhaltlich breit gefächerten Book on Demand-Veröffentlichungen.

Zu Jahresbeginn 2017 wurde die maschinelle Erschließung auch auf digitale Zeitschriftenartikel ausgeweitet, die zurzeit allerdings noch nicht in die Deutsche Nationalbibliografie aufgenommen sind. Das Importverfahren für E-Journals wurde Anfang 2016 gestartet. Allein 2016 wurden etwa 675 000 Zeitschriftenartikel in den Bestand der DNB integriert. Beginnend mit den Zeitschriften des Springer-Verlages reichert die DNB jetzt erstmalig auch die einzelnen Artikel mit inhaltserschließenden Metadaten an. In Anbetracht der Mengen ist eine Erschließung periodisch erscheinender Netzpublikationen auf dieser Ebene nur durch die Anwendung maschineller Methoden leistbar.

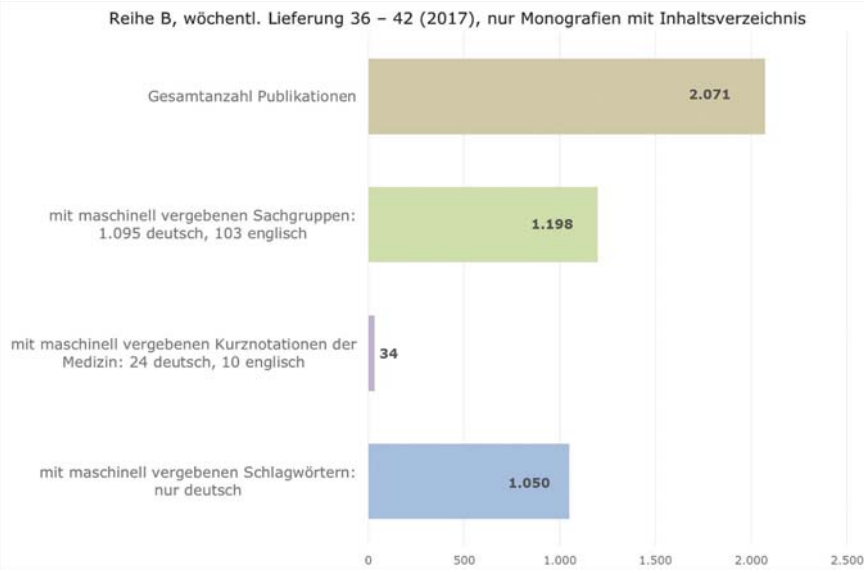


Abbildung 3: Anzahl der Publikationen mit maschineller Erschließung der wöchentlichen Lieferungen Nr. 36 – Nr. 42 (2017) der Reihe B (nur Monografien mit digitalisiertem Inhaltsverzeichnis).

Mit der Ausweitung der maschinellen Erschließung auf die gedruckten Monografien der Reihen B und H der Deutschen Nationalbibliografie wurde im September 2017 ein weiterer strategischer Meilenstein erreicht. Für die Literatursuche bedeutet dies, dass jetzt auch Hochschulschriften (Reihe H) und Publikationen, die außerhalb des Verlagsbuchhandels erscheinen (Reihe B), mit Schlagwörtern versehen werden (siehe Abbildung 3). Die DNB verzichtet für diese Reihen fortan auf die bisherige Tiefenerschließung mit der DDC. Diese soll für möglichst alle Sachgruppen

schrittweise durch DDC-Kurznotationen ersetzt werden, die maschinell vergeben werden können. Die Publikationen des Verlagsbuchhandels (Reihe A) werden weiterhin intellektuell erschlossen.

Für die maschinelle Analyse der physischen Medienwerke sollen perspektivisch alle vorhandenen digitalen Informationen herangezogen werden, beispielsweise parallele Online-Ausgaben, Inhaltsverzeichnisse oder Abstracts, Klappen- und Umschlagstexte. Zurzeit wird die Erschließung auf der Basis der digitalisierten Inhaltsverzeichnisse und der mitgelieferten bibliografischen Angaben durchgeführt. Aufgrund geringerer Textmengen und des teilweise niedrigen Informationsgehalts der Inhaltsverzeichnisse sind die Analysebedingungen allerdings häufig ungünstiger als bei den Netzpublikationen. Die maschinell vergebenen Sachgruppen werden bei den Reihen B und H daher durchgängig intellektuell überprüft.

### Kennzeichnung der Herkunft

Mit der Produktivnahme der Prozesse ist die Entscheidung verknüpft, die maschinell vergebenen Metadaten im Titeldatensatz zu verzeichnen, im Portal anzuzeigen, für das Retrieval zu nutzen und über die Datendienste auszuliefern. Auch die Erschließungsdaten der Zeitschriftenartikel stehen für die Literatursuche im Portal der DNB zur Verfügung und können über die Datendienste bezogen werden. Die DNB hat ihre internen Datenstrukturen angepasst, um die Herkunft und Vertrauenswürdigkeit der maschinell vergebenen Metadaten dokumentieren zu können. In der Katalogisierungsdatenbank werden sie jeweils zusammen mit dem Tagesdatum, dem Namen der Konfiguration sowie dem Konfidenzwert, einem Schätzwert zur Informationsgüte, verzeichnet. Bei der Anzeige im DNB-Portal werden die maschinell vergebenen DDC-Kurznotationen und Schlagwörter gekennzeichnet (siehe Abbildung 4).

Außerdem wurde das Datenaustauschformat MARC 21 angepasst, um Informationen zur Datenherkunft standardisiert mit ausliefern zu können. Hier steht die Angabe »maschinell gebildet« im MARC-Feld 883. Für die Sachgruppen und Schlagwörter wird die Information bereits exportiert. Das Verfahren für die Kurznotationen ist in Vorbereitung, ebenso wie eine Implementierung für den Linked-Data-Service der DNB.

### Qualität und Kontrolle

Neben einer täglichen Kontrolle des Prozessablaufs werden fachliche Überprüfungen in Form von Stichproben durchgeführt. Eine Auswahl der maschinell analysierten Publikationen wird somit zusätzlich auch intellektuell klassifiziert und beschlagwortet. Alle im Erschließungsprozess entstehenden Metadaten werden im Titeldatensatz dokumentiert. Immer dann, wenn auch intellektuell vergebene Metadaten vorhanden sind, werden diese für die Portal- und Datendienste bevorzugt genutzt.

Im Rahmen des Qualitätsmanagements wird die Güte der maschinellen Klassifikation durch Vergleich der Erschließungsdaten maschineller und intellektueller Herkunft statistisch ausgewertet. Hierfür werden auch die gegebenenfalls vorhandenen Daten paralleler Ausgaben mit herangezogen. Über den Zeitraum der letzten fünf Jahre hat die DNB etwa 18 Prozent der maschinell vergebenen Sachgruppen der Reihe O betrachtet. Dabei stimmten in 76 Prozent der Vergleichsfälle die maschinell und die intellektuell vergebenen Sachgruppen überein. In einigen Fächern wurde dieser Durchschnittswert sogar deutlich übertroffen, zum Beispiel im Recht mit 92 Prozent und in der Medizin mit 87 Prozent identischen Einordnungen. Allerdings funktioniert die maschinelle Klassifikation insbesondere für Fächer mit geringem Literaturaufkommen noch nicht zufriedenstellend, weil das Trainingsmaterial für die Lernprozesse nicht ausreicht. Ein Beispiel hierfür ist die Geschichte Südamerikas.

Bei der Überprüfung der Schlagwörter wird demgegenüber eine differenzierte Einzelbetrachtung durchgeführt, ob ein maschinell vergebenes Schlagwort für das Retrieval der Publikation nützlich oder ob der Suchbegriff falsch ist. Die

| | |
|---|---|
| *Link* | http://d-nb.info/1140134612 |
| *Titel* | Belastungen im Medizinstudium : eine Längsschnittuntersuchung zur Depressivität Medizinstudierender der Universität Jena |
| *Person(en)* | Hof, Katharina (Verfasser) |
| *Hochschulschrift* | Dissertation, Friedrich-Schiller-Universität Jena, 2017 |
| *Sprache(n)* | Deutsch (ger) |
| *Schlagwörter* | **Depressivität*** ; **Jena*** ; **Resilienz*** ; **Medizinstudium*** ; **Student*** ; **Längsschnittuntersuchung*** (*maschinell ermittelt) |
| *DDC-Notation* | **610.7** [maschinell ermittelte Kurznotation] |
| *Sachgruppe(n)* | **610** Medizin, Gesundheit ; **370** Erziehung, Schul- und Bildungswesen |
| *Inhaltsverzeichnis* | http://d-nb.info/1140134612/04 |

Abbildung 4: Titelanzeige einer maschinell erschlossenen Publikation der Reihe H mit Schlagwörtern, DDC-Kurznotation der Medizin sowie intellektuell geprüften DDC-Sachgruppen im DNB-Portal.

Bewertungen werden statistisch ausgewertet, um das Qualitätsniveau systematisch zu beobachten und Trends zu erkennen. Für den Jahrgang 2016 der Reihe O haben die Auswertungen zu dem Ergebnis geführt, dass etwa 78 Prozent der Schlagwörter in die Bewertungskategorien »sehr nützlich« bis hin zu »wenig nützlich« eingeordnet wurden, etwa 22 Prozent der maschinell vergebenen Schlagwörter sind falsch. Unbefriedigende Ergebnisse werden insbesondere immer dann erzielt, wenn die inhaltlich wesentlichen Begriffe noch nicht in der GND vorhanden sind. Deshalb ist die GND-Pflege ein wichtiger Ansatzpunkt für die Verbesserung der maschinellen Schlagwortvergabe.

### Verzahnung maschineller und intellektueller Erschließung

Die maschinellen Erschließungsverfahren arbeiten nicht fehlerfrei. Neben ungenauen und falschen Zuordnungen entsteht auch unnötiger Ballast. Aufgabe des Qualitätsmanagements ist es, die Fehlerquoten und ihre Auswirkungen auf den Datenbestand kritisch zu beobachten und bei Bedarf nachzusteuern. Ziel ist eine hohe Verlässlichkeit der Erschließungsdaten, unabhängig davon, ob sie intellektuell oder maschinell erzeugt wurden. Perspektivisch sollen intellektuelle und maschinelle Verfahren stärker miteinander verzahnt werden. Das Qualitätsmanagement dient der Steuerung und der Bewertung, welche Publikationsgruppen maschinell erschlossen werden können und welche Erschließungsleistungen intellektuell erbracht werden müssen.

Die bisherigen Erfahrungen zeigen die Bedeutung eines gut gepflegten Schlagwortvokabulars für die Qualität der Schlagwortvergabe. Zurzeit wird an Methoden gearbeitet, die auch solche Terme im Text als relevant erkennen, die bisher noch nicht im Schlagwortvokabular enthalten sind. Diese Begriffe sollen den GND-Redakteuren dann als neue Schlagwörter zur Einarbeitung in die Normdatei vorgeschlagen werden.

Der große Anteil englischsprachiger Netzpublikationen wird noch nicht mit computerlinguistischen Methoden beschlagwortet. Zurzeit wird daran gearbeitet, auch die Library of Congress Subject Headings (LCSH)[11] als Terminologie in das System einzubinden. Darüber hinaus soll eine Vernetzung mit anderen Datenressourcen wie DBpedia[12] oder YAGO[13] getestet werden. Crosskonkordanzen, beispielsweise zwischen LCSH und GND, werden als Option gesehen, gegebenenfalls auch mehrsprachige Sucheinstiege zu generieren.

Der breite Sammelauftrag der DNB stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse. So unterscheiden sich sprachliche Ausdrucksweisen nicht nur von Fachgebiet zu Fachgebiet, sondern oft auch innerhalb einer Fachdisziplin. Hinzu kommen die Unterschiede zwischen Wissenschaftssprache und Allgemeinsprache. Deshalb ist es für die Software schwierig, die inhaltlich wesentlichen Terme in einer Publikation immer richtig zu erkennen und einzuordnen. Es sind noch erhebliche Anstrengungen notwendig, die Fähigkeiten des Erschließungssystems weiter zu verbessern, beispielsweise durch Erweiterung und Kombination der Methoden.

Die Deutsche Nationalbibliothek hat außerdem damit begonnen, einen Rahmenplan für die Neugestaltung der Erschließungsumgebung zu entwickeln. Betrachtet werden alle Systemkomponenten und Dienste, die die Erschließungsarbeit und Metadatenverwaltung betreffen. Das schließt den Aufbau einer modernen Infrastruktur zur Pflege und Verwaltung der GND sowie Assistenzfunktionen zur Unterstützung der Qualitätssicherung mit ein.

Mit dieser Strategie stellt sich die DNB den Herausforderungen der voranschreitenden Digitalisierung und den Anforderungen an das Suchen und Finden der Publikationen in den heutigen Informationssystemen.

**Elisabeth Mödden** studierte Bauingenieurwesen an der Technischen Universität Braunschweig und absolvierte an der Universitäts- und Landesbibliothek Darmstadt das Bibliotheksreferendariat. Seit 2007 arbeitet sie an der Deutschen Nationalbibliothek, zunächst als Fachreferentin für Informatik und Technik, seit 2014 leitet sie das standortübergreifende Referat Automatische Erschließungsverfahren, Netzpublikationen. Kontakt: e.moedden@dnb.de

**Christa Schöning-Walter** ist Diplom-Informatikerin und hat eine Stabsstelle im Fachbereich Erwerbung und Erschließung der Deutschen Nationalbibliothek. Von 2009 bis 2014 leitete sie mit dem Projekt PETRUS (Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek) den Aufbau der maschinellen Erschließungsverfahren. Zuvor war sie als wissenschaftliche Referentin beim Deutschen Zentrum für Luft- und Raumfahrt (DLR) in der Projektträgerschaft tätig. Kontakt: c.schoening@dnb.de

**Sandro Uhlmann** arbeitet seit 2007 in der Deutschen Nationalbibliothek, zunächst in der Abteilung Inhaltserschließung, mittlerweile im Referat Automatische Erschließungsverfahren, Netzpublikationen mit dem Schwerpunkt Maschinelle Beschlagwortung. Kontakt: s.uhlmann@dnb.de

American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

# Predicted Geodetic Reference System for Baghdad City with Aided International Terrestrial Reference Frame (ITRF08)

Eng. Salman N Dawood[a*], Asst. Prof. Dr. Mustafa T. Mustafa[b], Asst. Prof. Dr. Abdulhaq Hadi Abed Ali[c]

[a]*MSc student in survey engineering; Technical College of Baghdad*

[b]*Head of Building and Construction Technical College*

[c]*Head of Highway and Transportation Dept Faculty of Engineering; Mustansiriyah University*

**Abstract**

Historically, the mean Earth ellipsoid is obtained by fitting an ellipsoid of revolution to the geoid. Such an ellipsoid, however, does not necessarily best fit the physical surface of the earth due to the existence of topography outside the geoid. When the distance between geoid and ellipsoid is as low as possible, GPS measurements are accurate because it depends on the measurement on the ellipsoid surface. The ellipsoid is defined as the shape produced by rotating an ellipse about one of its axes, which is a more correct definition mathematically ( Deng, X., 2013). An ellipsoid satisfying the condition that the deviations between the geoid and ellipsoid (in a local sense) are minimized. In this paper, presented a purely geometrically defined earth ellipsoid that best fits the physical surface of the Earth so that the resulting geoid undulation (N) attains minimum in Baghdad city. Using orthometric height (H) and ellipsoid height (h), the size, shape, position of such an Earth ellipsoid have been from observation GPS, methods DGPS with (ITRF).

---

* Corresponding author.

The establishment of a new geodetic reference frame of Baghdad city based on ITRF system which is compatible with space positioning techniques. One of the fundamental objectives of geodesy is to accurately define positions of points on the surface of the Earth. It is important and necessary to establish an accurate geodetic reference frame for measurements and computations of points on the surface of the earth. Recently, this has seen the advancement of technology of using GPS for determination of a three-dimensional geocentric reference system.

***Key words:*** Earth ellipsoid; Geoid; Best-fitting; Orthometric Height; Ellipsoid height; Geoid Undulation; ITRF; CORS.

## 1. Introduction

One of the most fundamental tasks of geodesy is to determine the figure of the earth. The shape of the earth is approximated by the mean earth ellipsoid, an ellipsoid of revolution that best fits the global geoid. The geodetic data from State Commission of Survey by the (POLESERVICE) company in 1979. (POLESERVICE) company in Iraq worked a vertical reference system and horizontal reference system in the 1970s. The horizontal reference was based on ellipsoid clark1880. In Iraq, the company worked to bring the elliptical closer to Geoid, establish land control networks. Also reducing N values as low as possible, for accurate coordinates and producing maps. The vertical reference was based on sea level, which was considered the point of Faw zero, in the reference system of Iraq. The production of maps in Iraq was based on Reference system Clarke 1880. Distance N between the geoid and best fitting ellipsoid is called geoid undulation and can be computed from N=h-H. The ellipsoid is as close as possible to the shape of the geoid within the boundaries of the city. It changes the re position of the global ellipsoid WGS84 to fit the city of Baghdad to be a local geodetic reference system for map production and accurate survey works.

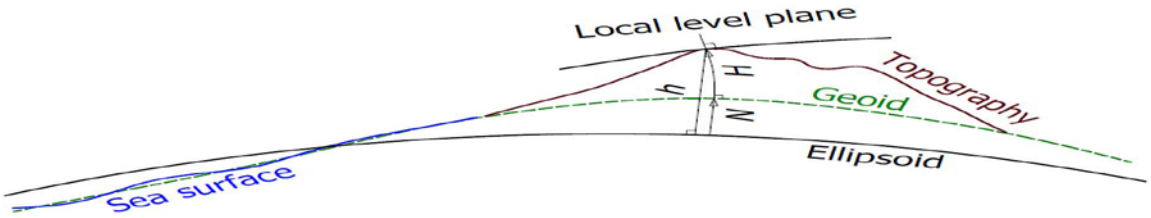


**Figure 1:** Surfaces, Orthometric Heights, Normal Heights and N values [7].

**Ellipsoidal heights (h)** are the heights of the location, normal (at right angles) to the reference ellipsoid [5].

**Orthometric heights (H)** are heights for a given position on the earth's surface above the geoid or MSL following a curved plumb line (Figure 1) [5].

The height difference between the geoid and a specific ellipsoid is known as the geoidal height or the geoid-ellipsoid separation (N ). The 'N' value is the ellipsoid height minus the orthometric height (Equations 1 and 2) [5].

$$N = h - H \quad (1)$$

$$H = h - N \quad (2)$$

GNSS natively uses ellipsoidal heights for data processing. These are converted to MSL elevations using an orthometric height from (POLESERVICE) report which computes N values for any given location in Baghdad.

**2. Best fits ellipsoid and geoid**

The spheroid is a very good approximation to the geoid, but there are significant differences. If a spheroid is formed which best fits the geoid. The height of the geoid above the spheroid is known as the geoid-spheroid separation, or often just the separation, and is usually given the symbol N. This may be a positive or a negative quantity. Every country or region that chooses an elliptical is always suitable for it, where the geoid is fits to the ellipsoid. As shown in Figure 2 [1].

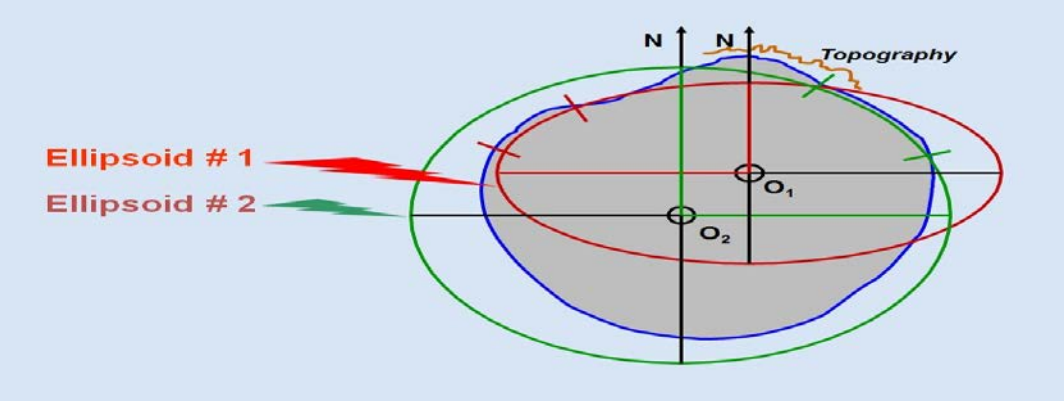


**Figure 2:** variable two ellipsoid with two zone [1].

## 3. Experimental Work And Data Collection

The horizontal and vertical measurements for the selected points of the local geodetic network have been taken from the State Commission of Survey and Mayoralty of Baghdad which were previously determined by POLESERVICE company.

These selected points have been observed utilizing GPS (Leica type GR15) by DGPS methods over various time variable as shown table (1).

To conduct a local correction, CORS stations near to the observed points were considered and a GPS LEICA type GR15 was utilized based on LEICA Geo Office program. The calculation of the new three-dimensional coordinates system had been carried out using the WGS84 ellipsoid system and based on Baghdad CORS station. Which adopted on observe the points Baghdad (ISBA) of the CORS stations and may be post-processing the CORS stations as shown figure (4). This can also assure accuracy in computing the ellipsoid height and orthometric height from the difference between the geoid and ellipsoid; this difference is called the geoid undulation (N) (Banerjee, P., G. R 1999). as shown table (2).

These programs can be also used along with Baghdad CORS station to correct the observed points to determine the local geodetic coordinates based on Ellipsoid WGS84. The value of the geoid undulation at all observed points have been computed. In addition, the observed vertical points are with WGS84 coordinates($\phi$, $\lambda$,h) and ITRF 2008. Corrected coordinates were based on CORS Baghdad (ISBA); include computed geographic coordinates as shown table (3). The seven unknown parameters (Translation, Rotation, Scale) were then computed between the WGS84 system and the new geodetic system of Baghdad as shown table (4), table 5). Furthermore, the next calculation step included determining the value of the geoid undulation for all the observed points between orthomatric height (from POLESERVICE report) and ellipsoid height (from GPS based on WGS84) as shown in Tables (2). And then developing acomparison between orthomatric height and ellipsoid WGS84 height with local geodetic (new geoid undulation)as shown Table (3) .

## 4. Research methodology

This study can be conducted through the following steps :

1- Investigate and specify points for this study.

2- Collect data for the geodetic coordinates of the studied area (Baghdad), provided by the ellipsoid (Clarke 1880) of Horizontal and Vertical National Geodetic Network and Iraqi Network, Geospatial Reference System (IGRS).

3- Observed the specified points and the nearest one to the CORS stations by of points by DGPS static observation with post-processing the data by (LEICA Geo Office) software program version 8.1 to get the accurate geodetic coordinates with(WGS84) reference.

4- Verity the value N (geoid undulation or geoid separation) for all points between (POLESERVICE) and WGS84.

5- Calculate the rate of geoid undulation and subtract the average value from all points.

6- Verity all points observation with new geoid undulation (N) ) between Geoid (POLESERVICE) and new geodetic reference for Baghdad city.

**Table 1:** Geodetic Coordinates for observed Horizontal and Vertical points National Geodetic Network (Ellipsoid WGS84 /ITRF2008).

| No. | Name of point | Latitude($\phi$) | Longitude($\lambda$) | Ellipsoid Height **(m)** |
|---|---|---|---|---|
| 1 | 20108 | 33º 14' 8.9474" | 44º 29' 23.0654" | 55.233 |
| 2 | 20081 | 33º 22' 23.7309" | 44º 31' 21.9874" | 58.933 |
| 3 | 20073 | 33º 24' 14.7503" | 44º 17' 52.3269" | 39.247 |
| 4 | 20080 | 33º 20' 14.0360" | 44º 23' 31.1269" | 84.838 |
| 5 | 527003 | 33º 21' 53.9122" | 44º 15' 0.7972 | 32.295 |
| 6 | 525504 | 33º 18' 19.7270" | 44º 16' 35.0450" | 34.141 |
| 7 | 517301 | 33º 25' 40.9742" | 44º 25' 1.83023" | 38.451 |
| 8 | 523002 | 33º 17' 35.1032" | 44º 27' 15.9473" | 32.333 |
| 9 | 524001 | 33° 13' 36.02694" | 44° 22' 27.86923" | 32.180 |
| 10 | 511701 | 33º 21' 37.9443" | 44º 21' 18.3075" | 35.637 |
| 11 | 32/2 | 33° 22' 39.4728" | 44° 19' 05.8638" | 36.797 |

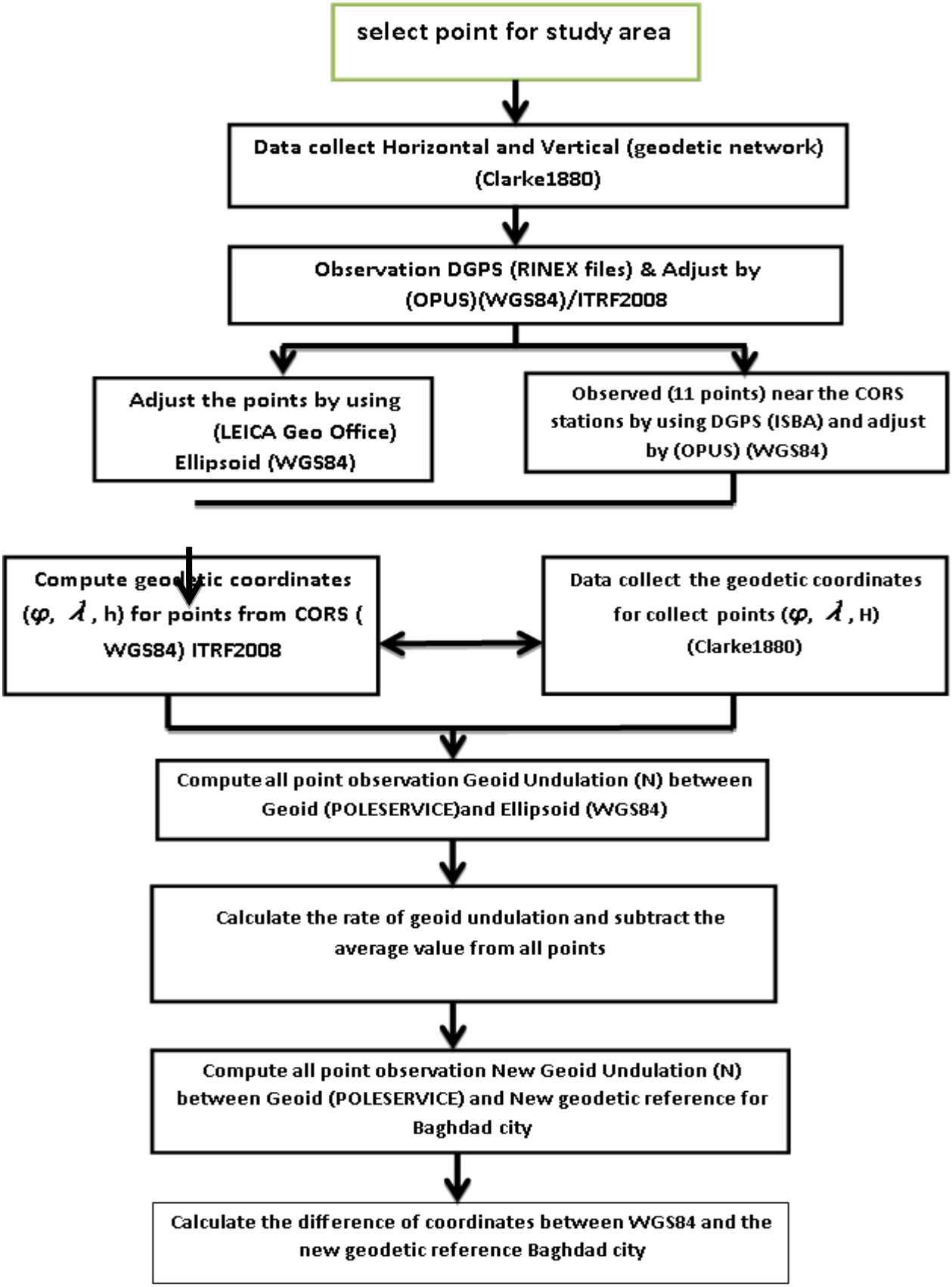


**Figure 3:** Schematic diagram of the methodology.

**Table 2:** The Geoid undulation between orthometric height (POLESREVICE) and Ellipsoid height

(WGS84/ITRF2008)

| No. | Name of point | Ellipsoid Height (m) | Orthometric Height (m) | Geoid Undulation (m) |
|---|---|---|---|---|
| 1 | 20108 | 55.233 | 57.20 | -1.967 |
| 2 | 20081 | 58.933 | 60.70 | -1.767 |
| 3 | 20073 | 39.241 | 40.20 | -0.959 |
| 4 | 20080 | 84.838 | 85.90 | -1.062 |
| 5 | 527003 | 32.295 | 33.314 | -1.019 |
| 6 | 525504 | 34.141 | 35.358 | -1.217 |
| 7 | 517301 | 38.451 | 40.077 | -1.626 |
| 8 | 523002 | 32.333 | 33.966 | -1.633 |
| 9 | 524001 | 30.389 | 32.180 | -1.791 |
| 10 | 511701 | 35.637 | 36.554 | -0.917 |
| 11 | 32/2 | 35.916 | 36.797 | -0.881 |
| | | | Average | 1.349 |

**Table 3:** The Geoid undulation between Orthometric height (POLESREVICE) and New Ellipsoid height (Baghdad city)

| No. | Name of point | New Ellipsoid Height (m) | Orthometric ELEV. (m) | New Geoid Undulation (m) |
|---|---|---|---|---|
| 1 | 20108 | 55.851 | 57.2 | -0.618 |
| 2 | 20081 | 59.351 | 60.7 | -0.418 |
| 3 | 20073 | 38.851 | 40.2 | 0.39 |
| 4 | 20080 | 84.551 | 85.9 | 0.287 |
| 5 | 527003 | 31.965 | 33.314 | 0.33 |
| 6 | 525504 | 34.009 | 35.358 | 0.132 |
| 7 | 517301 | 38.728 | 40.077 | -0.277 |
| 8 | 523002 | 32.617 | 33.966 | -0.284 |
| 9 | 524001 | 30.831 | 32.18 | -0.442 |
| 10 | 511701 | 35.205 | 36.554 | 0.432 |
| 11 | 32/2 | 35.448 | 36.797 | 0.468 |

**Table 4:** Geographic Coordinates for selected Horizontal and Vertical points Local Observed by (New Geodetic

Reference system).

| N | Description | Latitude ($\phi$) | Longitude ($\lambda$) | New Geoid Undulation(m) | Δ Latitude | Δ Longitude |
|---|---|---|---|---|---|---|
| 1 | 20108 | 33º 14' 08.9474"N | 44º 29' 23.06540"E | -0.618 | 0º 0' 0.02"N | 0º 0' 0.01"E |
| 2 | 20081 | 33º 22' 23.7309"N | 44º 31' 21.98741"E | -0.418 | 0º 0' 0.02"N | 0º 0' 0.01"E |
| 3 | 20073 | 33º 24' 14.7503"N | 44º 17' 52.32691"E | 0.39 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 4 | 20080 | 33º 20' 14.0360"N | 44º 23' 31.12691"E | 0.287 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 5 | 527003 | 33° 21' 53.89554" N | 44° 15' 00.78797"E | 0.33 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 6 | 525504 | 33° 18' 19.71011" N | 44° 16' 35.03610"E | 0.132 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 7 | 517301 | 33° 25' 40.93261" N | 44° 25' 01.82503" E | -0.277 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 8 | 523002 | 33° 17' 35.08228" N | 44° 27' 15.94200" E | -0.284 | 0º 0' 0.02"N | 0º 0' 0.01"E |
| 9 | 524001 | 33° 13' 36.02694" N | 44° 22' 27.86923" E | -0.442 | 0º 0' 0.01"N | 0º 0' 0.01"E |
| 10 | 511701 | 33° 21' 37.93224" N | 44° 21' 18.29184" E | 0.432 | 0º 0' 0.01"N | 0º 0' 0.02"E |
| 11 | 32/2 | 33° 22' 39.47279" N | 44° 19' 05.86379" E | 0.468 | 0º 0' 0.01"N | 0º 0' 0.01"E |

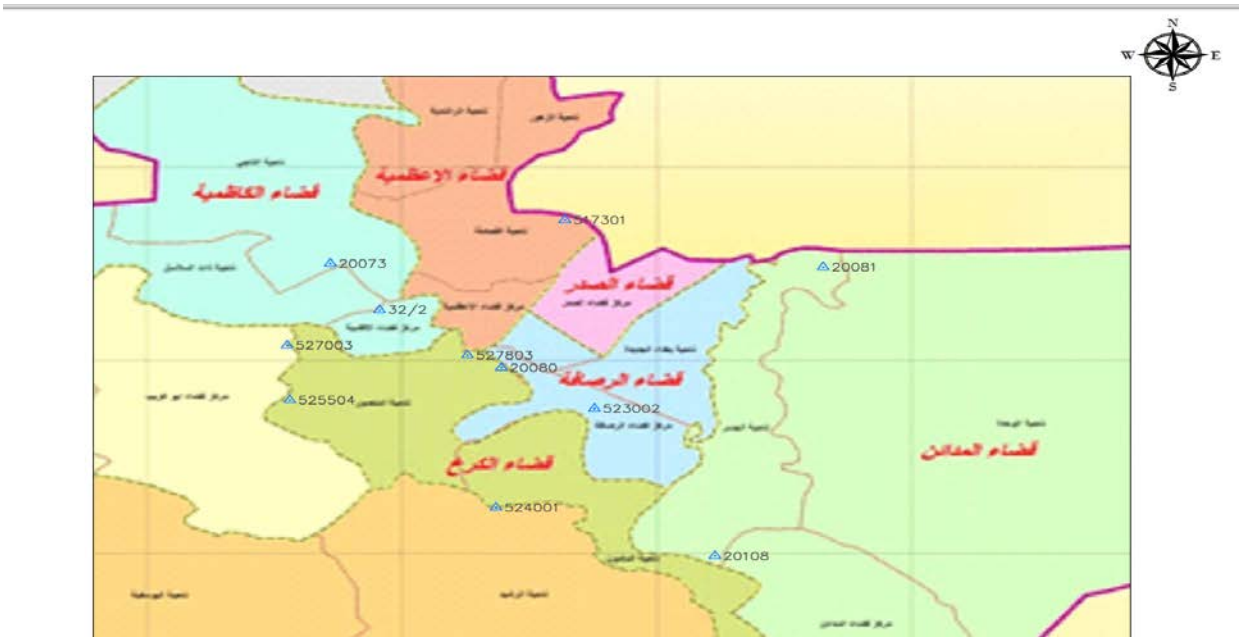


**Figure 5:** Location of the study area (Baghdad city) [9].


**Figure 4:** point correction with CORS Baghdad.

**Table 5:** Geographic coordinate WGS84 and coordinate new geodetic reference system and difference between two systems.

| | | WGS84/ITRF2008 | | | New geodetic reference system Baghdad city | | | | |
|---|---|---|---|---|---|---|---|---|---|
| N | Description | Latitude | Longitude | h | Latitude | Longitude | h | Δ Latitude | Δ Longitude |
| 1 | 20108 | 33° 14' 8.9674"N | 44° 29' 23.07154"E | -1.967 | 33° 14' 08.9474"N | 44° 29' 23.06540"E | -0.618 | 0° 0' 0.02"N | 0° 0' 0.01"E |
| 2 | 20081 | 33° 22' 23.71115"N | 44° 31' 21.99846"E | -1.767 | 33° 22' 23.7309"N | 44° 31' 21.98741"E | -0.418 | 0° 0' 0.02"N | 0° 0' 0.01"E |
| 3 | 20073 | 33° 24' 14.76191"N | 44° 17' 52.33808"E | -0.959 | 33° 24' 14.7503"N | 44° 17' 52.32691"E | 0.39 | 0° 0' 0.01"N | 0° 0' 0.01"E |
| 4 | 20080 | 33° 20' 14.03710"N | 44° 23' 31.1369"E | -1.062 | 33° 20' 14.0360"N | 44° 23' 31.12691"E | 0.287 | 0° 0' 0.01"N | 0° 0' 0.01"E |
| 5 | 527003 | 33° 21' 53.9122"N | 44° 15' 0.7972"E | -1.019 | 33° 21' 53.87554" N | 44° 15' 00.78797"E | 0.33 | 0° 0' 0.04"N | 0° 0' 0.01"E |
| 6 | 525504 | 33° 18' 19.7270"N | 44° 16' 35.0450"E | -1.217 | 33° 18' 19.69011" N | 44° 16' 35.03610"E | 0.132 | 0° 0' 0.04"N | 0° 0' 0.01"E |
| 7 | 517301 | 33° 25' 40.9742"N | 44° 25' 1.83023"E | -1.626 | 33° 25' 40.93261" N | 44° 25' 01.82503" E | -0.277 | 0° 0' 0.04"N | 0° 0' 0.01"E |
| 8 | 523002 | 33° 17' 35.1032"N | 44° 27' 15.9473"E | -1.633 | 33° 17' 35.06228" N | 44° 27' 15.94200" E | -0.284 | 0° 0' 0.04"N | 0° 0' 0.01"E |
| 9 | 524001 | 33° 13' 36.03786"N | 44° 22' 27.87714" E | -1.791 | 33° 13' 36.02694" N | 44° 22' 27.86923" E | -0.442 | 0° 0' 0.01"N | 0° 0' 0.01"E |
| 10 | 511701 | 33° 21' 37.9443"N | 44° 21' 18.3075"E | -0.917 | 33° 21' 37.93224" N | 44° 21' 18.29184" E | 0.432 | 0° 0' 0.01"N | 0° 0' 0.02"E |
| 11 | 32/2 | 33° 22' 39.4728"N | 44° 19' 05.8638"E | -0.881 | 33° 22' 39.47279" N | 44° 19' 05.86379" E | 0.468 | 0° 0' 0.01"N | 0° 0' 0.01"E |

## 5. Conclusions

- Geodetic surveying specialists often adopt a national or local geodetic reference system for a country or a specific region. Producing geographic maps based on this reference can lead to consistent and accurate results especially when the adopted reference surface simulates as much as possible the shape of the earth in the region.

In the works of geodesy, the size of the area needs to be surveyed should be carefully selected such that the distances among the geodetic points do not exceed 35 Km. The reason is to minimize the errors in the survey computations and map productions[9].

- The value of N, which represents the difference between the geoid production by (POLESERVICE) and the ellipsoid (WGS84) reference surfaces, ranges from -1 m to -2 m in the city of Baghdad. This difference is relatively large for the city of Baghdad.

This difference can negatively affects the accuracy of ground points coordinates and hence their real locations on the ground. Consequently, it can lead to significant inaccuracy in surveying applications such as the production of maps and the construction of infrastructure projects.

- The difference between geoid and ellipsoid (WGS84) was reduced by suggesting a new coordinate system in order to obtain a higher precision.

The new coordinates were linked to the satellite coordinates for the purpose of updating the maps based on adopted on an reference stations such as Baghdad station or any other stations in the country. That is because it depends on the global reference WGS84 where the center of WGS84 is the same center of gravity (geoid).

- In accordance with the development of the Geodesy, each area should be based on a base maps. In Iraq, the universal reference system Clarke 1880 was modified in the 1980s of the last century to be the Iraqi system that is nationally adopted in all surveying works – it was referred to as Karbala79. Regarding the vertical reference, the mean sea level (geoid) is still adopted [9].

## 6. Recommendations

The establishment of a geodetic reference to be utilized for assuring precise surveys requires at first the identification and determination of reference points that are approved by governmental institutions.

- Furthermore, the establishment of a geodetic reference should be complemented with adopting a proper vertical reference. It is important that this vertical system be continuously updated to match real earth (geoid). Several factors could contribute in changing the shape of the earth over the years including the tide effects, most of the countries updating their systems every ten years.
- In this research, due to the absence of an updating vertical reference, the vertical reference previously proposed by POLESERVICE in 1979 was adopted[9].
- On the other side, the correction of the coordinates of the observed points is highly recommended to be based on reference stations that are as much as possible close to these observed points; that is in order to ensure high accuracy.
- Furthermore, modern software that are compatible with GPS observations should be used in the process of coordinates correction. The exact paths of the satellites have to uploaded and entered in order to coordinates with high accuracy.
- Another important aspects that are relevant to the accuracy of observation process, includes the avoidance of high building, the selection of adequate time, and also the examining of operation status of the nearby CORS station.
- Finally yet importantly, it is recommended that the methodology proposed in this research is nationally adopted and implemented to create new Iraqi reference system.

## 7. Discussion

In the 1970s a completely new geodetic network covering all of Iraq was established by the Polish State Enterprise for Geodesy and Cartography GEOKART, operating as part of a larger foreign aid organization named Polservice. This was a traditional astro-geodetic control network consisting of 2778 points with an average inter-point distance of 15 km. The horizontal network was supplemented by a new spirit-levelled vertical network, tied to two tide gauges at the port of Al-Faw. Gravity observations were made along all precise levelling lines. The Fundamental Point of the horizontal network was moved from Nahrwan to Karbala. The

change

in coordinates of any point in Iraq is significant: in the Baghdad area the apparent ‘shift’ between Nahrwan and Karbala 1979 coordinates exceeds 400 metres. Although most of the documentation is still available, many of the control point monuments have been damaged.

Karbala 1979 was used in conjunction with the UTM system of map projections, as well as with a dedicated TM projection, known as the Iraq National Grid whose area of use is all of onshore Iraq.

The geodetic positioning is important for precise approximation of terrestrial and inertial reference system required to define local horizontal geodetic datum. Although the large number of the available local horizontal geodetic datums, which exceeds several hundreds, the number of the local horizontal datums for any selected area are significantly in decrease continuously. In order to accepted and implement a global geodetic datum, within a specified area, it is required to transform it between the geodetic datums.

The Iraqi national reference system, known as karbala-1979, was set by Geokart–Poland company in the late1970s. The system is a Clark1880 ellipsoid that is transformed to fit Iraq.

To accomplish the conversion from global to regional, the local geodetic datum and The World Geodetic System 1984 (WGS84) coordinates are both required at one or more sites within the local selected study area, so that the shift between the two datums can be computed. Satellite stations positioned within WGS84, for known local geodetic datum coordinates, are the basic components to convert between any local geodetic datum and the WGS 84 datum [7].

**References**

[1] Brunner, F.K. ed., 2013. Advances in Positioning and Reference Frames: IAG Scientific Assembly Rio de Janeiro, Brazil, September 3–9, 1997 (Vol. 118). Springer Science & Business Media.

[2] Beutler, G. and Rummel, R., 2012. Scientific rationale and development of the global geodetic observing system. In Geodesy for Planet Earth (pp. 987-993). Springer, Berlin, Heidelberg.

[3] Bossler, John D. "Datums and geodetic systems." Manual of Geospatial Science and Technology (2004):

16-26.

[4] Banerjee, P., G. R. Foulger, and C. P. Dabral. "Geoid undulation modelling and interpretation at Ladak, NW Himalaya using GPS and levelling data." Journal of Geodesy 73.2 (1999): 79-86.

[5] Charles, D.G. and Paul, R.W., 2008. Elementary Surveying: An Introduction to Geomatics.

[6] Deng, X., 2013. Geodesy–Introduction to Geodetic Datum and Geodetic Systems.

[7] Jekeli, C., 2006. Geometric reference systems in geodesy. Division of Geodesy and Geospatial Science, School of Earth Sciences, Ohio State University, 25.

[8] National Oceanic and Atmospheric Administration.

[9] Authority General Surveying.