Data audit for technical evaluation of **Spider (Arachnida, Araneae) fauna of the lowland part of the Balkhash-Alakol basin (SE Kazakhstan): an integrated assessment** Resource links:
https://www.gbif.org/dataset/92cc4b3a-97f2-4aa3-b260-883ba061c1a9
https://www.gbif.org/dataset/791d91fe-ce27-4044-8a3d-a2a8b203f979 the Darwin Core archives **dwca-almaty_literature-v1.2.zip** and **dwca-almaty_collection-v1.1.zip** were downloaded on 2025-09-22 from:
http://gbif.ru:8080/ipt/archive.do?r=almaty_literature
http://gbif.ru:8080/ipt/archive.do?r=almaty_collection

Maxim Shashkov (m.shashkov@pensoft.net)

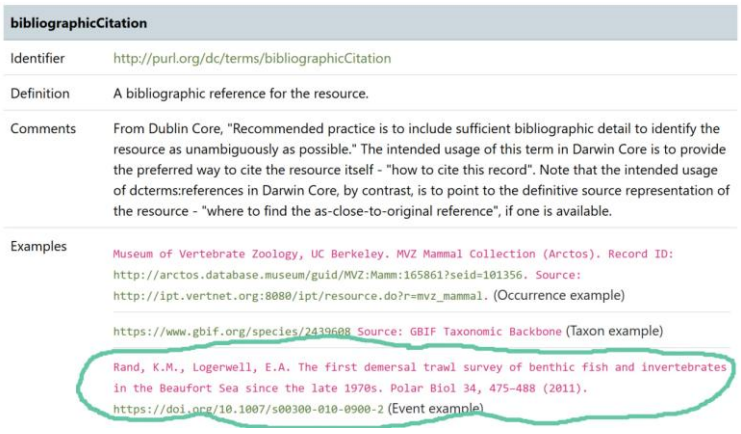
"part1" first:

1. There are only three allowed invisible characters for dataset tables: horizontal tab (u0009), whitespace (u0020), and linefeed (u000a). Any others can cause errors during processing. There is a soft hyphen (SHY, u00ad, c2 ad) in 1 records line 1401, field 31 (marker with {HERE}):
   Taldy-Korgan, Dzhambul Distr., Alma-Ata - Geor{HERE}gievka highway, ca. 8 km W of Targan

fixed

2. The ***bibliographicCitation*** field is intended to provide the preferred way to cite the resource itself. Source of the primary data should be specified using the ***associatedReferences*** field.

As far as the data are from literature, its citation is required. Reference on the published articles can be used “to cite this records”, so is allowed in
https://dwc.tdwg.org/terms/#dcterms:bibliographicCitation
and moreover, provided as one of valid examples:



bibliographicCitation	
Identifier	http://purl.org/dc/terms/bibliographicCitation
Definition	A bibliographic reference for the resource.
Comments	From Dublin Core, "Recommended practice is to include sufficient bibliographic detail to identify the resource as unambiguously as possible." The intended usage of this term in Darwin Core is to provide the preferred way to cite the resource itself - "how to cite this record". Note that the intended usage of dcterms:references in Darwin Core, by contrast, is to point to the definitive source representation of the resource - "where to find the as-close-to-original reference", if one is available.
Examples	Museum of Vertebrate Zoology, UC Berkeley. MVZ Mammal Collection (Arctos). Record ID: http://arctos.database.museum/guid/MVZ:Mamm:165861?seid=101356. Source: http://ipt.vertnet.org:8080/ipt/resource.do?r=mvz_mammal. (Occurrence example) https://www.gbif.org/species/2439608 Source: GBIF Taxonomic Backbone (Taxon example) Rand, K.M., Logerwell, E.A. The first demersal trawl survey of benthic fish and invertebrates in the Beaufort Sea since the late 1970s. Polar Biol 34, 475-488 (2011). https://doi.org/10.1007/s00300-010-0900-2 (Event example)

3. What the difference between the ***institutionID*** and ***institutionCode*** field?

   You have an institution identifier specified within the GRSciColl registry: https://scientific-collections.gbif.org/institution/7e82dc97-c81e-4361-845f-4338170452b2 Please use it for one of those fields.
   In addition, you specify one of the alternative codes of the institute as the code of collection (IZRK), while no collection is provided in the GRSciColl registry for you institution: https://registry.gbif.org/institution/7e82dc97-c81e-4361-845f-4338170452b2/collection.
   A proper way would be to register the collection you described in the dataset through GRSciColl and give the newly defined acronym both in the dataset and manuscript.

Many thanks, the values of institutionID field were updated by the proper URL link.
Values of the institutionCode field remain the same as far as DwC allows to fill it by “the name of institution” not only by its acronym.
Collection registration through GRSciColl was applied and available by URL: https://registry.gbif.org/collection/de8790ba-69fd-4934-aa46-5921e4fb3110.

4. Please specify the ***datasetName*** field according to the dataset title. (for both datasets)

fixed

5. The most appropriate ***basisOfRecord*** for the data extracted from literature is "MaterialCitation": https://dwc.tdwg.org/list/#dwc_MaterialCitation

fixed

6. Not clear usage of the ***informationWithheld*** field. Usually this field is intended for some details that authors retain unpublished for some reasons.

Not all data extracted from the processed literature belong the lowland part of the region so shouldn't be presented in the dataset, technically. However, a lot of efforts was applied to extract these, and literature-based data are minority on GBIF, so it is valuable data which better not to hide. Therefore, the ***informationWithheld*** field serves here as separator which occurrences comes from plain part and accounted in plain fauna ({"occurrence":"plain"}), and which ones comes from mountain part and excluded from analysis ({"occurrence":"mountain"}). Hiding these data is the worst way. Publishing them separately doesn’t seem good as well due to splitting occurrences from the same article into different datasets.

Noted in the manuscript

7. "NA" is not a valid value. If the real value is unknown, just leave it blank. ***informationWithheld***, ***individualCount***, ***sex***, ***lifeStage***, ***samplingProtocol***, ***habitat*** ***eventRemarks***, ***minimumElevationInMeters***, ***maximumElevationInMeters*** and others within the both datasets.

fixed

8. Please unify punctuation within the ***occurrenceRemarks*** field:

deposited in Zoological Museum of the Moscow State University (ZMMU), Moscow, Russia (Ta-4692)
deposited in Zoological Museum of the Moscow State University (ZMMU), Moscow, Russia, (Ta-4693)
deposited in Zoological Museum of the Moscow State University (ZMMU), Moscow, Russia (Ta-4695)

unified

9. Please use vertical bar - " | " to separate multiple values: ***recordedBy***, ***georeferencedBy***, ***identifiedBy*** (both datasets)

Tarabaev Ch.K., Zlatanova A.A., Tyschenko V.P. ->
Tarabaev Ch.K. | Zlatanova A.A. | Tyschenko V.P.

for dataset 2 as well

By the way, who is "V.L." ?

fixed

V.L. is abbreviated collector originally provided by authors (P. 39). It is impossible to get more information.

10. The ***occurrenceStatus*** can not be "present" if the ***individualCount*** is zero.

Zero here was wrong value. Correct value (NA) is inserted. Now there is no discordance between columns.

11. The final ***eventDate*** can not be earlier than the initial:

"1995-04-29/04-27"

The similar for dataset 2: "2024-12-05/2023-11-12"

fixed

12. The earliest ***eventDate*** is "1829". Please verify against the dataset metadata and manuscript text.

fixed

13. Please unify spelling within the ***habitat*** field:

2 sandy semi-desert
1 sandy semidesert

1 stony mountain steppe with rocks
2 stony mountain steppe, with rocks

unified

14. Geographic coordinates specified with excessive precision. Please round up to 5th decimal place which corresponds to meters on the ground. Moreover, to specify coordinates with an uncertainty of 100 m, it is enough to provide coordinates with 4 decimal places, 1000 m with 3, and 10000 with 2.

For dataset "part 2" as well.

fixed

However, too much rounding conflicts with explicitly defined coordinateUncertaintyInMeters field. Point-radius method clearly reflects the precision of the point. The rounding shifts the

center of the circle so extending of coordinateUncertaintyInMeters field may become necessary: + 1000/2 m in case after rounding to 3, + 10000/2 m in case of rounding to 2 etc.

15. Are "Spassky S." and "Spassky S.A." the same person? in the ***identifiedBy*** field. Please verify and unify

Yes, the person is the same. We tried to save the values as close to the literature origin as possible here as well as in another fields. One article sets author (who obviously identified taxa) as Spassky S. when others as Spassky S.A. We would suggest remain it as is, because difference of these data and the original text can become an obstacle for machine learning in future.

16. Please unify spelling of the ***taxonRemarks*** field:

    29 sp. n.: holotype
    4 sp.n.: holotype

    5 sp. n.: paratypes
    2 sp.n.: paratypes

unified, as well as “gen.n.”

17. There are many values with doubled whitespaces:

    Taldy-Korgan, Taldy-Korgan Distr., near Kospal
    Almaty (=Alma-Ata) Area, Kerbulak District, c. 25 km SSW of Basshi (=Baschi, Kalinino), c. 0.9 km NE of Kosbastau
    Almaty [= Alma-Ata] Region, Ili Distr., E slope of Karaoi Plateau, N vicinities of Kapchagai Town
    and others
    for dataset "part 2" as well

fixed

18. There is inconsistency between the ***scientificName*** and ***scientificNameAuthorship***:

    No. of records | scientificName | scientificNameAuthorship

    20 | Xysticus pseudocristatus Azarkina & Logunov, 2001 | (Clerck, 1757)
    6 | Xysticus pseudocristatus Azarkina & Logunov, 2001 | Azarkina & Logunov, 2001

Taxon field schema is changed. Now ***scientificName*** contains originally provided taxa (including authorship if known) and ***acceptedNameUsage*** contains currently correct taxa (mistypes, synonymy, misidentification etc.)

following issues for dataset 2

19. Two field contain only NA values: ***institutionID*** and ***occurrenceRemarks***. If the real values are completely unknown, please remove these fields.

institutionID is filled properly, ***occurrenceRemarks*** is removed

20. "unknown" is not a valid value for the ***recordedBy***.

removed

21. There are inconsistencies between eventDate and the decomposed date: year, month, and day:

2023-05-28/2024-06-05	2023 5	28
2023-05-28/2024-06-05	2024 6	
2023-08-02/2024-07-25	2023 8	2
2023-08-02/2024-07-25	2024 7	25
2023-11-10/2024-08-06	2024	
2024-05-28/06-03	2024 5	28
2024-05-28/06-03	2024 6	3
2024-12-05/2023-11-12	2024	

If it is not possible to specify an exact distinct year, month, and/or day, just leave the corresponding field blank.

removed

22. Please unify the spelling of the ***scientificName***:

2 Allagelena gracilens (C. L. Koch, 1841)
5 Allagelena gracilens (C.L.Koch, 1841)

1 Drassyllus praeficus (L. Koch, 1866)
3 Drassyllus praeficus (L.Koch, 1866)
and others similar

unified

23. Please unify authorship spelling:

"C. L. Koch", "C. L. Koch"
"O. Pickard-Cambridge", "O.Pickard-Cambridge"
…

both for ***scientificName*** and ***scientificNameAuthorship*** fields and for dataset "part 1" as well.

unified

24. Please unify ***taxonRemarks*** field:

1 Similar to A. hui/zonsteini
1 Similar to A. hui/zonsteini.

unified

25. Please provide a field list for both datasets in the **Data resources** section of the manuscript.

done

26. Please clarify the ***Contacts*** section of the metadata for dataset "Part 1". Persons mentioned there were duplicated.

done